Penalty-Regulated Dynamics and Robust Learning Procedures in Games

Pierre Coucheney, Bruno Gaujal, Panayotis Mertikopoulos

PENALTY-REGULATED DYNAMICS AND ROBUST LEARNING PROCEDURES IN GAMES

PIERRE COUCHENEY, BRUNO GAUJAL, AND PANAYOTIS MERTIKOPOULOS

Abstract. Starting from a heuristic learning scheme for N-person games, we derive a new class of continuous-time learning dynamics consisting of a replicator-like drift adjusted by a penalty term that renders the boundary of the game's strategy space repelling. These penalty-regulated dynamics are equivalent to players keeping an exponentially discounted aggregate of their on-going payoffs and then using a smooth best response to pick an action based on these performance scores. Owing to this inherent duality, the proposed dynamics satisfy a variant of the folk theorem of evolutionary game theory and they converge to (arbitrarily precise) approximations of Nash equilibria in potential games. Motivated by applications to traffic engineering, we exploit this duality further to design a discrete-time, payoff-based learning algorithm which retains these convergence properties and only requires players to observe their in-game payoffs: moreover, the algorithm remains robust in the presence of stochastic perturbations and observation errors, and it does not require any synchronization between players.

1. Introduction

Owing to the computational complexity of Nash equilibria and related game-theoretic solution concepts, algorithms and processes for learning in games have received considerable attention over the last two decades. Such procedures can be divided into two broad categories, depending on whether they evolve in continuous or discrete time: the former class includes the numerous dynamics for learning and evolution (see e.g. Sandholm [34] for a recent survey), whereas the latter focuses on learning algorithms (such as fictitious play and its variants) for infinitely iterated games – for an overview, see Fudenberg and Levine [11] and references therein.

A key challenge in these endeavors is that it is often unreasonable to assume that players can monitor the strategies of their opponents – or even calculate the payoffs of actions that they did not play. As a result, much of the literature on learning in games focuses on payoff-based schemes that only require players to observe the stream of their *in-game* payoffs: for instance, the regret-matching procedure of Hart and Mas-Colell [12, 13] converges to the set of correlated equilibria (in an empirical, time-average sense), whereas the trial-and-error process of Young [42] guarantees with high probability that players will spend a large proportion of their time near a pure Nash equilibrium (provided that such an equilibrium exists).

The authors are greatly indebted to the associate editor and two anonymous referees for their insightful suggestions, and to M. Bravo and R. Cominetti for many helpful discussions and remarks.

This work was supported by the European Commission in the framework of the FP7 Network of Excellence in Wireless COMmunications NEWCOM# (contract no. 318306) and the French National Research Agency (ANR) project NETLEARN (grant no. ANR-13-INFR-004).

In this paper, we focus on a reinforcement learning framework in which players score their actions over time based on their observed payoffs and they then employ a smooth best response map (such as logit choice) to determine their actions at the next instance of play. Learning mechanisms of this kind have been investigated in continuous time by Börgers and Sarin [5], Rustichini [33], Hopkins [17], Hopkins and Posch [18], Tuyls et al. [39] and many others: Hopkins [17] in particular showed that in 2-player games, the continuous-time dynamics that correspond to this learning process may be seen as a variant of the replicator dynamics with an extra penaly term that keeps players from attaining the boundary of the game's strategy space (see also Hopkins and Posch [18]). On the other hand, from a discrete-time viewpoint, Leslie and Collins [24] used a Q-learning approach to establish the convergence of the resulting learning algorithm in 2-player games under minimal information assumptions; in a similar vein, Cominetti et al. [9] and Bravo [7] took a moving-average approach for scoring actions in general N-player games and provided sufficient convergence conditions for the resulting dynamics. Interestingly, in all these cases, when the learning process converges, it converges to a so-called quantal response equilibrium (QRE) which is a fixed point of a *perturbed* best response correspondence – as opposed to the standard notion of Nash equilibrium which is a fixed point of the *unperturbed* best response map; see e.g. McKelvey and Palfrey [27].

Discrete-time processes of this kind are usually analyzed by means of stochastic approximation (SA) techniques that are used to compare the long-term behavior of the discrete-time process to the corresponding mean-field dynamics in continuous time – for a comprehensive introduction to the subject, see e.g. Benaïm [3] and Borkar [6]. Indeed, there are several conditions which guarantee that a discrete-time process and its continuous counterpart both converge to the same sets, so continuous dynamics are usually derived as the limit of (possibly random) discrete-time processes – cf. the aforementioned works by Leslie and Collins [24], Cominetti et al. [9] and Bravo [7].

Contrary to this approach, we descend from the continuous to the discrete and we develop two different learning processes from the same dynamical system (the actual algorithm depends crucially on whether we look at the evolution of the players' strategies or the performance scores of their actions). Accordingly, the first contribution of our paper is to derive a class of *penalty-regulated* game dynamics consisting of a replicator-like drift plus a penalty term that keeps players from approaching the boundary of the state space. These dynamics are equivalent to players scoring their actions by comparing their exponentially discounted cumulative payoffs over time and then using a smooth best response to pick an action; as such, the class of penalty-regulated dynamics that we consider constitutes the strategy-space counterpart of the Q-learning dynamics of Leslie and Collins [24], Hopkins [17] and Tuyls et al. [39]. Thanks to this link to the replicator dynamics, the dynamics converge to quantal response equilibria in potential games, and we also establish a variant of the folk theorem of evolutionary game theory (Hofbauer and Sigmund [15]). In particular, we show the dynamics' stability and convergence depends crucially on the discount factor used by the players to score their strategies over time: in the undiscounted case, strict Nash equilibria are the only attracting states, just as in the replicator equation; on the other hand, for positive discount factors, only QRE that are close to strict equilibria remain asymptotically stable.

The second contribution of our paper concerns the implementation of these dynamics as a learning algorithm with the following desirable properties:

(1) The learning process is *distributed* and *stateless*: players update their strategies using only their observed in-game payoffs and no further knowledge.
(2) The algorithm retains its convergence properties even if the players' observations are subject to stochastic perturbations and observation errors (or even if they are not up-to-date).
(3) Updates need not be synchronized – there is no need for a global timer used by all players.

These desiderata are key for the design of robust, decentralized optimization protocols in network and traffic engineering, but they also pose significant challenges. Nonetheless, by combining the long-term properties of the continuous-time dynamics with stochastic approximation techniques, we show that players converge to arbitrarily precise approximations of strict Nash equilibria whenever the game admits a potential function (cf. Theorem 4.4 and Proposition 4.6). Thus, thanks to the congestion characterization of such games (Monderer and Shapley [30]), we obtain a distributed robust optimization method for a wide class of engineering problems, ranging from traffic routing to wireless communications – see e.g. Altman et al. [1], Mertikopoulos et al. [28] and references therein.

1.1. **Paper outline and structure.** After a few preliminaries, our analysis proper begins in Section 2 where we introduce our cumulative reinforcement learning scheme and derive the associated penalty-regulated dynamics. Owing to the duality between the players' mixed strategies and the performance scores of their actions (measured by an exponentially discounted aggregate of past payoffs), we obtain two equivalent formulations: the score-based equation (PRL) and the strategy-based dynamics (PD). In Section 3, we exploit this interplay to derive the long-term convergence properties of the dynamics; finally, Section 4 is devoted to the discretization of the dynamics (PRL) and (PD) and their implementation as bona fide learning algorithms.

1.2. **Notational conventions.** If $\mathcal{S} = \{s_\alpha\}_{\alpha=0}^n$ is a finite set, the real space spanned by $\mathcal{S}$ will be denoted by $\mathbb{R}^\mathcal{S}$ and its canonical basis by $\{e_s\}_{s\in\mathcal{S}}$. To avoid drowning in a morass of indices, we will make no distinction between $s \in \mathcal{S}$ and the corresponding basis vector e_s of $\mathbb{R}^\mathcal{S}$, and we will frequently use the index α to refer interchangeably to either s_α or e_α (writing e.g. x_α instead of x_{s_α}). Likewise, if $\{\mathcal{S}_k\}_{k\in\mathcal{K}}$ is a finite family of finite sets indexed by $k \in \mathcal{K}$, we will use the shorthands $(\alpha_k;\alpha_{-k})$ for the tuple $(\dots,\alpha_{k-1},\alpha_k,\alpha_{k+1},\dots) \in \prod_k \mathcal{S}_k$ and we will write $\sum_\alpha^k$ instead of $\sum_{\alpha\in\mathcal{S}_k}$.

The set $\Delta(\mathcal{S})$ of probability measures on $\mathcal{S}$ will be identified with the unit n-dimensional simplex $\Delta(\mathcal{S}) \equiv \{x \in \mathbb{R}^\mathcal{S} : \sum_\alpha x_\alpha = 1 \text{ and } x_\alpha \geq 0\}$ of $\mathbb{R}^\mathcal{S}$. Finally, regarding players and their actions, we will follow the original convention of Nash and employ Latin indices $(k,\ell,\dots)$ for players, while keeping Greek ones $(\alpha,\beta,\dots)$ for their actions (pure strategies); also, unless otherwise mentioned, we will use $\alpha,\beta,\dots$, for indices that start at 0, and $\mu,\nu,\dots$, for those which start at 1.

1.3. **Definitions from game theory.** A *finite game* $\mathfrak{G} \equiv \mathfrak{G}(\mathcal{N}, \mathcal{A}, u)$ will be a tuple consisting of *a*) a finite set of *players* $\mathcal{N} = \{1, \dots, N\}$; *b*) a finite set $\mathcal{A}_k$ of *actions* (or *pure strategies*) for each player $k \in \mathcal{N}$; and *c*) the players' *payoff functions* $u_k \colon \mathcal{A} \to \mathbb{R}$, where $\mathcal{A} \equiv \prod_k \mathcal{A}_k$ denotes the game's *action space*, i.e. the set of all *action profiles* $(\alpha_1, \dots, \alpha_N)$, $\alpha_k \in \mathcal{A}_k$. A *restriction* of $\mathfrak{G}$ will then be a game $\mathfrak{G}' \equiv \mathfrak{G}'(\mathcal{N}, \mathcal{A}', u')$ with the same players as $\mathfrak{G}$, each with a subset $\mathcal{A}'_k \subseteq \mathcal{A}_k$ of their original actions, and with payoff functions $u'_k \equiv u_k|_{\mathcal{A}'}$ suitably restricted to the reduced action space $\mathcal{A}' = \prod_k \mathcal{A}'_k$ of $\mathfrak{G}'$.

Of course, players can mix their actions by taking probability distributions $x_k = (x_{k\alpha})_{\alpha \in \mathcal{A}_k} \in \Delta(\mathcal{A}_k)$ over their action sets $\mathcal{A}_k$. In that case, their expected payoffs will be

$$u_k(x) = \sum\nolimits_{\alpha_1}^{1} \cdots \sum\nolimits_{\alpha_N}^{N} u_k(\alpha_1, \dots, \alpha_N)\, x_{1,\alpha_1} \cdots x_{N,\alpha_N}, \tag{1.1}$$

where $x = (x_1, \dots, x_N)$ denotes the players' *strategy profile* and $u_k(\alpha_1, \dots, \alpha_N)$ is the payoff to player k in the (pure) action profile $(\alpha_1, \dots, \alpha_N) \in \mathcal{A}$;[1] more explicitly, if player k plays the pure strategy $\alpha \in \mathcal{A}_k$, we will use the notation $u_{k\alpha}(x) \equiv u_k(\alpha; x_{-k}) = u_k(x_1, \dots, \alpha, \dots, x_N)$. In this mixed context, the *strategy space* of player k will be the simplex $\mathcal{X}_k \equiv \Delta(\mathcal{A}_k)$ while the strategy space of the game will be the convex polytope $\mathcal{X} \equiv \prod_k \mathcal{X}_k$. Together with the players' (expected) payoff functions $u_k \colon \mathcal{X} \to \mathbb{R}$, the tuple $(\mathcal{N}, \mathcal{X}, u)$ will be called the *mixed extension* of $\mathfrak{G}$ and it will also be denoted by $\mathfrak{G}$ (relying on context to resolve any ambiguities).

The most prominent solution concept in game theory is that of Nash equilibrium (NE) which characterizes profiles that are resilient against unilateral deviations; formally, $q \in \mathcal{X}$ will be a *Nash equilibrium* of $\mathfrak{G}$ when

$$u_k(x_k; q_{-k}) \leq u_k(q) \quad \text{for all } x_k \in \mathcal{X}_k \text{ and for all } k \in \mathcal{N}. \tag{NE}$$

In particular, if (NE) is strict for all $x_k \in \mathcal{X}_k \setminus \{q_k\}$, $k \in \mathcal{N}$, q will be called a *strict* Nash equilibrium; finally, a *restricted equilibrium* of $\mathfrak{G}$ will be a Nash equilibrium of a restriction $\mathfrak{G}'$ of $\mathfrak{G}$.

An especially relevant class of finite games is obtained when the players' payoff functions satisfy the *potential property*:

$$u_{k\alpha}(x) - u_{k\beta}(x) = U(\alpha; x_{-k}) - U(\beta; x_{-k}) \tag{1.2}$$

for some (necessarily) multilinear function $U \colon \mathcal{X} \to \mathbb{R}$. When this is the case, the game will be called a *potential game with potential function* U, and as is well known, the pure Nash equilibria of $\mathfrak{G}$ will be precisely the vertices of $\mathcal{X}$ that are local maximizers of U (Monderer and Shapley [30]).

2. Reinforcement learning and penalty-regulated dynamics

Our goal in this section will be to derive a class of learning dynamics based on the following reinforcement premise: agents keep a long-term "performance score" for each of their actions and they then use a smooth best response to map these scores to strategies and continue playing. Accordingly, our analysis will comprise two components:

(1) The *assessment stage* (Section 2.1) describes the precise way with which players aggregate past payoff information in order to update their actions' performance scores.

[1] Recall that we will be using α for both elements $\alpha \in \mathcal{A}_k$ and basis vectors $e_\alpha \in \Delta(\mathcal{A}_k)$, so there is no clash of notation between payoffs to pure and mixed strategies.

(2) The *choice stage* (Section 2.2) then details how these scores are used to select a mixed strategy.

For simplicity, we will work here in continuous time and we will assume that players can observe (or otherwise calculate) the payoffs of all their actions in a given strategy profile; the descent from continuous to discrete time and the effect of imperfect information will be explored in Section 4.

2.1. **The assessment stage: aggregation of past information.** The aggregation scheme that we will consider is the familiar exponential discounting model:

$$y_{k\alpha}(t) = \int_0^t \lambda^{t-s} u_{k\alpha}(x(s))\, ds, \tag{2.1}$$

where $\lambda \in (0,\infty)$ is the model's discount rate, $x(s) \in \mathcal{X}$ is the players' strategy profile at time s and we are assuming for the moment that the model is initially unbiased, i.e. $y(0) = 0$. Clearly then:

(1) For $\lambda \in (0,1)$ the model assigns exponentially more weight to more recent observations.
(2) If $\lambda = 1$ all past instances are treated uniformly – e.g. as in Rustichini [33], Hofbauer et al. [16], Sorin [37], Mertikopoulos and Moustakas [29] and many others.
(3) For $\lambda > 1$, the scheme (2.1) instead assigns exponentially more weight to older instances.

With this in mind, differentiating (2.1) readily yields

$$\dot{y}_{k\alpha} = u_{k\alpha} - T y_{k\alpha}, \tag{2.2}$$

where

$$T \equiv \log(1/\lambda) \tag{2.3}$$

represents the *discount rate* of the performance assessement scheme (2.1). In tune with our previous discussion, the standard exponential discounting regime $\lambda \in (0,1)$ corresponds to positive $T > 0$, a discount rate of 0 means that past information is not penalized in favor of more recent observations, while $T < 0$ means that past observations are reinforced in favor of more recent ones.

Remark 1. Leslie and Collins [24] and Tuyls et al. [39] examined the aggregation scheme (2.2) from a quite different viewpoint, namely as the continuous-time limit of the Q-learning estimator

$$y_{k\alpha}(n+1) = y_{k\alpha}(n) + \gamma_{n+1}\left(u_{k\alpha}(x(n)) - y_{k\alpha}(n)\right) \times \frac{\mathbb{1}(\alpha_k(n+1) = \alpha)}{\mathbb{P}\left(\alpha_k(n+1) = \alpha \mid \mathcal{F}_n\right)}, \tag{2.4}$$

where $\mathbb{1}$ and $\mathbb{P}$ denote respectively the indicator and probability of player k choosing $\alpha \in \mathcal{A}_k$ at time $n+1$ given the history $\mathcal{F}_n$ of the process up to time n, while γ_n is a variable step-size with $\sum_n \gamma_n = +\infty$ and $\sum_n \gamma_n^2 < +\infty$ (see also Fudenberg and Levine [11]). The exact interplay between (2.2) and (2.4) will be explored in detail in Section 4; for now, we simply note that (2.2) can be interpreted both as a model of discounting past information and also as a moving Q-average.

Remark 2. We should also note here the relation between (2.4) and the moving average estimator of Cominetti et al. [9] that omits the factor $\mathbb{P}(\alpha_k(n+1) = \alpha \mid \mathcal{F}_n)$ (or the similar estimator of Bravo [7] which has a state-dependent step size). As a result of this difference, the mean-field dynamics of Cominetti et al. [9] are scaled by $x_{k\alpha}$, leading to the adjusted dynamics $\dot{y}_{k\alpha} = x_{k\alpha}(u_{k\alpha} - y_{k\alpha})$. Given this difference in form, there is essentially no overlap between our results and those of Cominetti et al. [9], but we will endeavor to draw analogies with their results wherever possible.

2.2. **The choice stage: smooth best responses.** Having established the way that agents evaluate their strategies' performance over time, we now turn to mapping these assessment scores to mixed strategies $x \in \mathcal{X}$. To that end, a natural choice would be for each agent to pick the strategy with the highest score via the mapping

$$y_k \mapsto \arg\max_{x_k \in \mathcal{X}_k} \sum_\beta^k x_{k\beta} y_{k\beta} \tag{2.5}$$

Nevertheless, this "best response" approach carries several problems: First, if two scores $y_{k\alpha}$ and $y_{k\beta}$ happen to be equal (e.g. if there are payoff ties), (2.5) becomes a multi-valued mapping which requires a tie-breaking rule to be resolved (and is theoretically quite cumbersome to boot). Additionally, such a practice could lead to completely discontinuous trajectories of play in continuous time – for instance, if the payoffs $u_{k\alpha}$ are driven by an additive white Gaussian noise process, as is commonly the case in information-theoretic applications of game theory; see e.g. Altman et al. [1]. Finally, since best responding generically leads to pure strategies, such a process precludes convergence of strategies to non-pure equilibria in finite games.

To circumvent these obstacles, we will replace the arg max operator with the regularized variant

$$Q_k(y_k) = \arg\max_{x_k \in \mathcal{X}_k} \left\{ \sum_\beta^k x_{k\beta} y_{k\beta} - h_k(x_k) \right\}, \tag{2.6}$$

where $h_k \colon \mathcal{X}_k \to \mathbb{R}$ is a smooth strongly convex function which acts as a *penalty* (or "control cost") to the maximization objective $\sum_\beta^k x_{k\beta} y_{k\beta}$ of player k.[2] Choice models of this type are known in the literature as *smooth best response maps* (or *quantal response functions*) and have seen extensive use in game-theoretic learning; for a comprehensive account, see e.g. van Damme [40], McKelvey and Palfrey [27], Fudenberg and Levine [11], Hofbauer and Sandholm [14], Sandholm [34] and references therein. Formally, following Alvarez et al. [2], we have:

Definition 2.1. Let $\mathcal{S}$ be a finite set and let $\Delta \equiv \Delta(\mathcal{S})$ be the unit simplex spanned by $\mathcal{S}$. We will say that $h \colon \Delta \to \mathbb{R} \cup \{+\infty\}$ is a *penalty function* on Δ if:

(1) h is finite except possibly on the relative boundary $\operatorname{bd}(\Delta)$ of Δ.
(2) h is continuous on Δ, smooth on $\operatorname{rel\,int}(\Delta)$, and $|dh(x)| \to +\infty$ when x converges to $\operatorname{bd}(\Delta)$.
(3) h is convex on Δ and strongly convex on $\operatorname{rel\,int}(\Delta)$.

[2]Note here that this penalty mechanism is different than the penalty imputed to past payoff observations in the performance assessment step (2.1): (2.1) discounts past instances of play whereas (2.6) discourages the player from choosing pure strategies. Despite this fundamental difference, these two processes end up being intertwined in the resulting learning scheme, so we will use the term "penalty" for both mechanisms, irrespective of origin.

We will also say that h is (*regularly*) *decomposable with kernel* θ if $h(x)$ can be written in the form:

$$h(x) = \sum\nolimits_{\beta\in\mathcal{S}} \theta(x_\beta) \tag{2.7}$$

where $\theta\colon [0,+1] \to \mathbb{R}\cup\{+\infty\}$ is a continuous function such that

a) θ is finite and smooth on $(0,1]$.
b) $\theta''(x) > 0$ for all $x \in (0,1]$.
c) $\lim_{x\to 0^+} \theta'(x) = -\infty$ and $\lim_{x\to 0^+} \theta'(x)/\theta''(x) = 0$.

In this context, the map $Q\colon \mathbb{R}^{\mathcal{S}} \to \Delta$ of (2.6) will be referred to as the *choice map* (or *smooth best response* or *quantal response function*) induced by h.

Given that (2.6) allows us to view $Q(\eta y) = \arg\max_{x\in\Delta}\{\sum_\beta x_\beta y_\beta - \eta^{-1}h(x)\}$ as a smooth approximation to the arg max operator in the limit $\eta \to \infty$ (i.e. when the penalty term becomes negligible), the choice stage of our learning process will consist precisely of the choice maps that are derived from penalty functions as above; for simplicity of presentation however, our analysis will mostly focus on the decomposable case.

In any event, Definition 2.1 will be central to our considerations, so some comments are in order:

Remark 1. The fact that choice maps are well-defined and single-valued is an immediate consequence of the convexity and boundary properties of h; the smoothness of Q then follows from standard arguments in convex analysis – see e.g. Chapter 26 in Rockafellar [32]. Moreover, the requirement $\lim_{x\to 0^+} \theta'(x)/\theta''(x) = 0$ of Definition 2.1 is just a safety net to ensure that penalty functions do not exhibit pathological traits near the boundary $\mathrm{bd}(\Delta)$ of Δ. As can be easily seen, this growth condition is satisfied by all of the example functions (2.8) below; in fact, to go beyond this natural requirement, θ'' must oscillate deeply and densely near 0.

Remark 2. Examples of penalty functions abound; some of the most prominent ones are:

1. The Gibbs entropy: $h(x) = \sum_\beta x_\beta \log x_\beta$. (2.8a)
2. The Tsallis entropy: $h(x) = (1-q)^{-1}\sum_\beta (x_\beta - x_\beta^q)$, $0 < q \leq 1$. (2.8b)
3. The Burg entropy: $h(x) = -\sum_\beta \log x_\beta$. (2.8c)

Strictly speaking, the Tsallis entropy is not well-defined for $q = 1$, but it approaches the standard Gibbs entropy as $q \to 1$, so we will use (2.8a) for $q = 1$ in that case.[3]

Example 1 (Logit choice). The most well-known example of a smooth best response is the so-called *logit map*

$$G_\alpha(y) = \frac{\exp(y_\alpha)}{\sum_\beta \exp(y_\beta)}, \tag{2.9}$$

which is generated by the Gibbs entropy $h(x) = \sum_\beta x_\beta \log x_\beta$ of (2.8a). For uses of this map in game-theoretic learning, see e.g. Cominetti et al. [9], Fudenberg and Levine [11], Hofbauer and Sandholm [14], Hofbauer et al. [16], Leslie and Collins [24], McFadden [26], Marsili et al. [25], Mertikopoulos and Moustakas [29], Rustichini [33], Sorin [37] and many others.

[3]Actually, entropies are concave in statistical physics and information theory, but this detail will not concern us here.

Remark 3. Interestingly, McKelvey and Palfrey [27] provide an alternative derivation of (2.9) as follows: assume first that the score vector y is subject to additive stochastic fluctuations of the form

$$\tilde{y}_\alpha = y_\alpha + \xi_\alpha, \tag{2.10}$$

where the ξ_α are independent Gumbel-distributed random variables with zero mean and scale parameter $\varepsilon > 0$ (amounting to a variance of $\varepsilon^2\pi^2/6$). It is then known that the *choice probability* $P_\alpha(y)$ of the α-th action (defined as the probability that α maximizes the perturbed variable $\tilde{y}_\alpha$) is just

$$P_\alpha(y) \equiv \mathbb{P}\left(\tilde{y}_\alpha = \max_\beta \tilde{y}_\beta\right) = G_\alpha(\varepsilon^{-1}y). \tag{2.11}$$

As a result, the logit map can be seen as either a smooth best response to the deterministic penalty function $h(x)$ or as a perturbed best response to the stochastic perturbation model (2.10); furthermore, both models approximate the ordinary best response correspondence when the relative magnitude of the perturbations approaches 0. In a more general context, Hofbauer and Sandholm [14] showed that this observation continues to hold even when the stochastic perturbations ξ_α are not Gumbel-distributed but follow an arbitrary probability law with a strictly positive and smooth density function: mutatis mutandis, the choice probabilities of a stochastic perturbation model of the form (2.10) can be interpreted as a smooth best response map induced by a deterministic penalty function in the sense of Definition 2.1.

2.3. **The dynamics of penalty-regulated learning.** Combining the results of the previous two sections, we will focus on the *penalty-regulated learning process*:

$$\begin{aligned} y_{k\alpha}(t) &= y_{k\alpha}(0)\, e^{-Tt} + \int_0^t e^{-T(t-s)} u_{k\alpha}(x(s))\, ds, \\ x_k(t) &= Q_k(y_k(t)), \end{aligned} \tag{PRL}$$

where $y_{k\alpha}(0)$ represents the initial bias of player k towards action $\alpha \in \mathcal{A}_k$,[4] T is the model's discount rate, and $Q_k \colon \mathbb{R}^{\mathcal{A}_k} \to \mathcal{X}_k$ is the smooth best response map of player k (induced in turn by some player-specific penalty function $h_k \colon \mathcal{X}_k \to \mathbb{R}$).

From an implementation perspective, the difficulty with (PRL) is twofold: First, it is not always practical to write the choice maps Q_k in a closed-form expression that the agents can use to update their strategies.[5] Furthermore, even when this is possible, (PRL) is a two-step, primal-dual process which does not allow agents to update their strategies directly. The rest of this section will thus be devoted to writing (PRL) as a continuous-time dynamical system on $\mathcal{X}$ that can be updated with minimal computation overhead.

To that end, we will focus on decomposable penalty functions of the form

$$h_k(x_k) = \sum\nolimits_\beta^k \theta_k(x_{k\beta}), \tag{2.12}$$

[4]The exponential decay of $y(0)$ is perhaps best explained by the differential formulation (2.2) for which $y(0)$ is an initial condition; in words, the player's initial bias simply dies out at the same rate as a payoff observation at $t = 0$.

[5]The case of the Gibbs entropy is a shining (but, ultimately, misleading) exception to the norm.

where the kernels θ_k, $k \in \mathcal{N}$, satisfy the convexity and steepness conditions of Definition 2.1.[6] In this context, the Karush–Kuhn–Tucker (KKT) conditions for the maximization problem (2.6) give

$$y_{k\alpha} - \theta_k'(x_{k\alpha}) = \zeta_k, \tag{2.13}$$

where ζ_k is the Lagrange multiplier for the equality constraint $\sum_\alpha^k x_{k\alpha} = 1$.[7] By differentiating, we then obtain:

$$\dot{y}_{k\alpha} - \theta_k''(x_{k\alpha})\dot{x}_{k\alpha} = \dot{\zeta}_k, \tag{2.14}$$

and hence, a little algebra yields:

$$\begin{aligned}
\dot{x}_{k\alpha} &= \frac{1}{\theta_k''(x_{k\alpha})}\big[\dot{y}_{k\alpha} - \dot{\zeta}_k\big] \\
&= \frac{1}{\theta_k''(x_{k\alpha})}\big[u_{k\alpha}(x) - Ty_{k\alpha} - \dot{\zeta}_k\big] \\
&= \frac{1}{\theta_k''(x_{k\alpha})}\big[u_{k\alpha}(x) - T\theta_k'(x_{k\alpha}) - \big(\dot{\zeta}_k + T\zeta_k\big)\big],
\end{aligned} \tag{2.15}$$

where the second equality follows from the definition of the penalty-regulated scheme (PRL) and the last one from the KKT equation (2.13). However, since $\sum_\alpha^k x_{k\alpha} = 1$, we must also have $\sum_\alpha^k \dot{x}_{k\alpha} = 0$; thus, summing (2.15) over $\alpha \in \mathcal{A}_k$ gives:

$$\dot{\zeta}_k + T\zeta_k = \Theta_k''(x_k)\sum\nolimits_\beta^k \frac{1}{\theta_k''(x_{k\beta})}\left[u_{k\beta}(x) - T\theta_k'(x_{k\beta})\right], \tag{2.16}$$

where Θ_k'' denotes the harmonic aggregate:[8]

$$\Theta_k''(x_k) = \left[\sum\nolimits_\beta^k 1/\theta_k''(x_{k\beta})\right]^{-1}. \tag{2.17}$$

In this way, by putting everything together, we finally obtain the *penalty-regulated dynamics*

$$\dot{x}_{k\alpha} = \frac{1}{\theta_k''(x_{k\alpha})}\left[u_{k\alpha}(x) - \Theta_k''(x_k)\sum\nolimits_\beta^k \frac{u_{k\beta}(x)}{\theta_k''(x_{k\beta})}\right] \tag{PD}$$

$$- \frac{T}{\theta_k''(x_{k\alpha})}\left[\theta_k'(x_{k\alpha}) - \Theta_k''(x_k)\sum\nolimits_\beta^k \frac{\theta_k'(x_{k\beta})}{\theta''(x_{k\beta})}\right], \tag{2.18}$$

Along with the aggregation-driven learning scheme (PRL), the dynamics (PD) will be the main focus of our paper, so some remarks and examples are in order:

Example 2 (The Replicator Dynamics). As a special case, the Gibbs kernel $\theta(x) = x\log x$ of (2.8a) leads to the *adjusted replicator equation*

$$\dot{x}_{k\alpha} = x_{k\alpha}\left[u_{k\alpha}(x) - \sum\nolimits_\beta^k x_{k\beta}u_{k\beta}(x)\right] - T\,x_{k\alpha}\left[\log x_{k\alpha} - \sum\nolimits_\beta^k x_{k\beta}\log x_{k\beta}\right]. \tag{RD$_T$}$$

[6] Non-decomposable h can be treated similarly but the end expression is more cumbersome so we will not present it.

[7] The complementary slackness multipliers for the inequality constraints $x_{k\alpha} \geq 0$ can be omitted because the steepness properties of θ_k ensure that the solution of (2.6) is attained in the interior of the simplex.

[8] Needless to say, Θ_h'' is not a second derivatives per se; we just use this notation for visual consistency.

As the name implies, when the discount rate T vanishes, (RD$_T$) freezes to the ordinary (asymmetric) replicator dynamics of Taylor and Jonker [38]:

$$\dot{x}_{k\alpha} = x_{k\alpha}\left[u_{k\alpha}(x) - \sum\nolimits_{\beta}^{k} x_{k\beta}u_{k\beta}(x)\right]. \tag{RD}$$

In this way, for $T = 0$, we recover the well-known equivalence between the replicator dynamics and exponential learning in continuous time – for a more detailed treatment, see e.g. Rustichini [33], Hofbauer et al. [16], Sorin [37] and Mertikopoulos and Moustakas [29].

Remark 4 (Links with existing dynamics). Leslie and Collins [24] derived a differential version of the penalty-regulated learning process (PRL) as the mean-field dynamics of the Q-learning estimator (2.4); independently, Tuyls et al. [39] obtained a variant of the strategy-space dynamics (PD) in the context of Q-learning in 2-player games. A version of (PD) for 2-player games also appeared in Hopkins [17] and Hopkins and Posch [18] as a perturbed reinforcement learning model; other than that however, the penalty-regulated dynamics (PD) appear to be new.

Interestingly, in terms of structure, the differential system (PD) consists of a replicator-like term driven by the game's payoffs, plus a game-independent adjustment term which reflects the penalty imputed to past payoffs. This highlights a certain structural similarity between (PD) and other classes of game dynamics with comparable correction mechanisms: for instance, in a stochastic setting, Itô's lemma leads to a "second order in space" correction in the stochastic replicator dynamics of Fudenberg and Harris [10], Cabrales [8], and Mertikopoulos and Moustakas [29]; likewise, such terms also appear in the "second order in time" approach of Laraki and Mertikopoulos [20, 21].

The reason for this similarity is that all these models are first defined in terms of a set of auxiliary variables: absolute population sizes in Fudenberg and Harris [10] and Cabrales [8], and payoff scores in Laraki and Mertikopoulos [20] and here. Differentiation of these "dual" variables with respect to time yields a replicator-like term (which carries the dependence on the game's payoffs) plus a game-independent adjustment which only depends on the relation between these "dual" variables and the players' mixed strategies (the system's "primal" variables).

Remark 5 (Well-posedness). Importantly, the dynamics (PD) are *well-posed* in the sense that they admit unique global solutions for every interior initial condition $x(0) \in \operatorname{rel\,int}(\mathcal{X})$. Since the vector field of (PD) is not Lipschitz, perhaps the easiest way to see this is by using the integral representation (PRL) of the dynamics: indeed, given that the payoff functions $u_{k\alpha}$ are Lipschitz and bounded, the scores $y_{k\alpha}(t)$ will remain finite for all $t \geq 0$, so interior solutions $x(t) = Q(y(t))$ of (PD) will be defined for all $t \geq 0$.

Moreover, even though the dynamics (PD) are technically defined only on the relative interior of the game's strategy space, the steepness and regularity requirements of Definition 2.1 allow us to extend the dynamics to the boundary $\operatorname{bd}(\mathcal{X})$ of $\mathcal{X}$ by continuity (i.e. by writing $1/\theta''(x_{k\alpha}) = \theta'(x_{k\alpha})/\theta''(x_{k\alpha}) = 0$ when $x_{k\alpha} = 0$). By doing just that, every subface $\mathcal{X}'$ of $\mathcal{X}$ will be forward invariant under (PD), so the class of penalty-regulated dynamics may be seen as a subclass of the imitative dynamics introduced by Björnerstedt and Weibull [4] (see also Weibull [41]).

Remark 6 (Sharpened choices). In addition to tuning the discount rate of the learning scheme (PRL), players can also sharpen their smooth best response model by replacing the choice stage (2.6) with

$$x_k = Q_k(\eta_k y_k) \tag{2.19}$$

for some $\eta_k > 0$. The choice parameters η_k may thus be viewed as (player-specific) *inverse temperatures*: as $\eta_k \to \infty$, the choice map of player k freezes down to the arg max operator, whereas in the limit $\eta_k \to 0$, player k will tend to mix actions uniformly, irrespectively of their performance scores.

In this context, the same reasoning as before leads to the rate-adjusted dynamics:

$$\dot{x}_{k\alpha} = \frac{\eta_k}{\theta_k''(x_{k\alpha})}\left[u_{k\alpha}(x) - \Theta_k''(x_k)\sum\nolimits_\beta^k \frac{u_{k\beta}(x)}{\theta_k''(x_{k\beta})}\right] \tag{PD$_\eta$}$$

$$- \frac{T}{\theta_k''(x_{k\alpha})}\left[\theta_k'(x_{k\alpha}) - \Theta_k''(x_k)\sum\nolimits_\beta^k \frac{\theta_k'(x_{k\beta})}{\theta''(x_{k\beta})}\right]. \tag{2.20}$$

We thus see that the parameters T and η play very different roles in (PD$_\eta$): the discount rate T affects only the game-independent penalty term of (PD$_\eta$) whereas η_k affects only the term which is driven by the game's payoffs.

3. Long-run rationality analysis

In this section, our aim will be to analyze the asymptotic properties of the penalty-regulated dynamics (PD) with respect to standard game-theoretic solution concepts. Thus, in conjunction with the notion of Nash equilibrium, we will also focus on the widely studied concept of *quantal response equilibria*:

Definition 3.1 (McKelvey and Palfrey [27])**.** Let $\mathfrak{G} \equiv \mathfrak{G}(\mathcal{N}, \mathcal{A}, u)$ be a finite game and assume that each player $k \in \mathcal{N}$ is endowed with a quantal response function $Q_k \colon \mathbb{R}^{\mathcal{A}_k} \to \mathcal{X}_k$ (cf. Definition 2.1). We will say that $q = (q_1, \dots, q_N) \in \mathcal{X}$ is a *quantal response equilibrium* (QRE) of $\mathfrak{G}$ with respect to Q (or a Q*-equilibrium* for short) when, for some $\rho \geq 0$ and for all $k \in \mathcal{N}$:

$$q_k = Q_k\left(\rho u_k(q)\right), \tag{QRE}$$

where $u_k(q) = (u_{k\alpha}(q))_{\alpha\in\mathcal{A}_k} \in \mathbb{R}^{\mathcal{A}_k}$ denotes here the payoff vector of player k. More generally, we will say that $q \in \mathcal{X}$ is a *restricted* QRE of $\mathfrak{G}$ if it is a QRE of some restriction $\mathfrak{G}'$ of $\mathfrak{G}$.

The scale parameter $\rho \geq 0$ will be called the *rationality level* of the QRE in question. Obviously, when $\rho = 0$, QRE have no ties to the game's payoffs; at the other end of the spectrum, when $\rho \to \infty$, quantal response functions approach best responses and the notion of a QRE approximates smoothly that of a Nash equilibrium. To see this in more detail, let $q^* \in \mathcal{X}$ be a Nash equilibrium of $\mathfrak{G}$, and let $\gamma \colon U \to \mathcal{X}$ be a smooth curve on $\mathcal{X}$ defined on a half-infinite interval of the form $U = [a, +\infty)$, $a \in \mathbb{R}$. We will then say that γ is a Q*-path to* q^* when $\gamma(\rho)$ is a Q-equilibrium of $\mathfrak{G}$ with rationality level ρ and $\lim_{\rho\to\infty}\gamma(\rho) = q^*$; in a similar vein, we will say that $q \in \mathcal{X}$ is a Q*-approximation* of q^* when q is itself a Q-equilibrium and there is a Q-path joining q to q^* (van Damme [40] uses the terminology *approachable*).

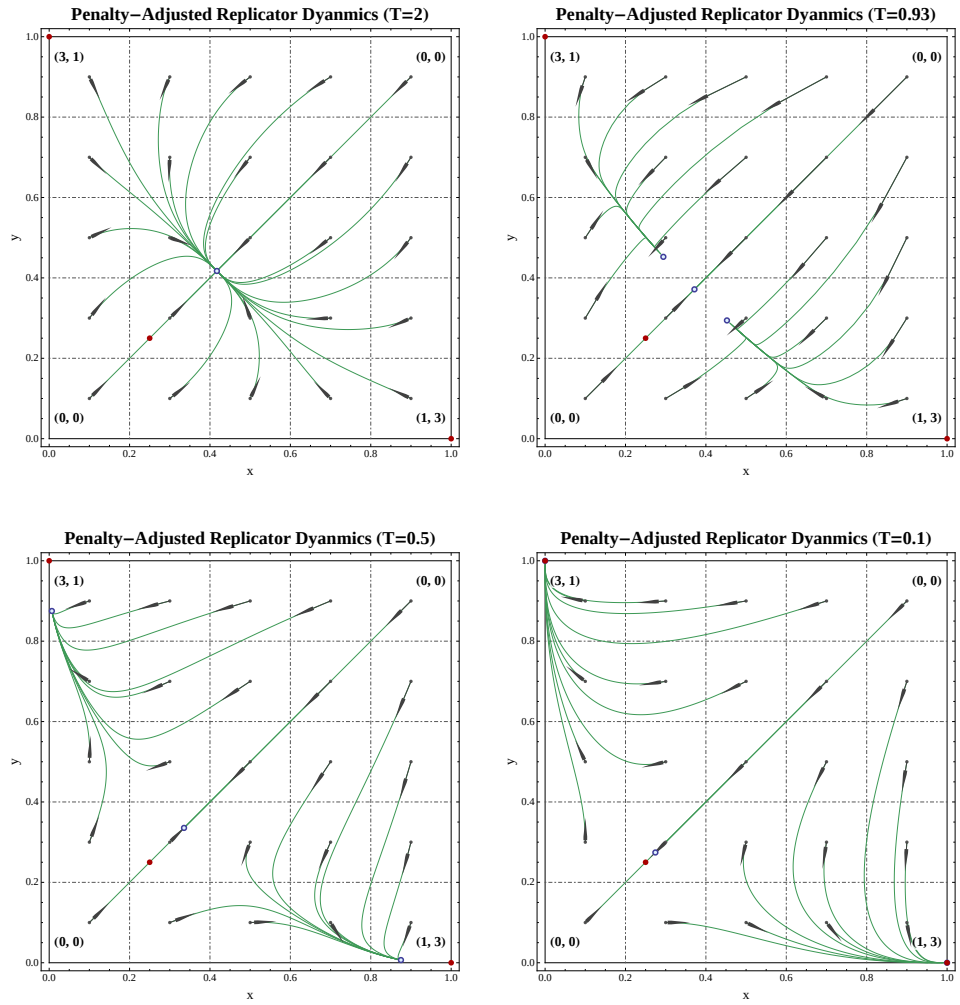


FIGURE 1. Phase portraits of the penalty-adjusted replicator dynamics (RD_T) in a 2×2 potential game (Nash equilibria are depicted in dark red and interior rest points in light/dark blue; for the game's payoffs, see the vertex labels). For high discount rates $T \gg 0$, the dynamics fail to keep track of the game's payoffs and their only rest point is a global attractor which approaches the barycenter of $\mathcal{X}$ as $T \to +\infty$ (corresponding to a QRE of very low rationality level). As the players' discount rate drops down to the critical value $T_c \approx 0.935$, the globally stable QRE becomes unstable and undergoes a supercritical pitchfork bifurcation (a phase transition) which results in the appearance of two asymptotically stable QRE that approach the strict Nash equilibria of the game as $T \to 0^+$.

Example 3. By far the most widely used specification of a QRE is the *logit equilibrium* which corresponds to the Gibbs choice map (2.9): in particular, we will say that $q \in \mathcal{X}$ is a logit equilibrium of $\mathfrak{G}$ when $q_{k\alpha} = \exp(\rho u_{k\alpha}(q))/\sum_\beta \exp(\rho u_{k\beta}(q))$ for all $\alpha \in \mathcal{A}_k$, $k \in \mathcal{N}$.

3.1. **Stability analysis.** We begin by linking the rest points of (PD) to the game's QRE:

Proposition 3.2. *Let $\mathfrak{G} \equiv \mathfrak{G}(\mathcal{N}, \mathcal{A}, u)$ be a finite game and assume that each player $k \in \mathcal{N}$ is endowed with a quantal response function $Q_k : \mathbb{R}^{\mathcal{A}_k} \to \mathcal{X}_k$. Then:*

(1) For $T > 0$, the rest points of the penalty-regulated dynamics (PD) coincide with the restricted QRE of $\mathfrak{G}$ with rationality level $\rho = 1/T$.
(2) For $T = 0$, the rest points of (PD) are the restricted Nash equilibria of $\mathfrak{G}$.

Proof. Proof. Since the proposition concerns restricted equilibria, it suffices to establish our assertion for interior rest points; given that the faces of $\mathcal{X}$ are forward-invariant under the dynamics (PD), the general claim follows by descending to an appropriate restriction $\mathfrak{G}'$ of $\mathfrak{G}$.

To wit, (PD) implies that any interior rest point $q \in \operatorname{rel\,int}(\mathcal{X})$ will have $u_{k\alpha}(q) - T\theta_k'(q_{k\alpha}) = u_{k\beta}(q) - T\theta_k'(q_{k\beta})$ for all $\alpha, \beta \in \mathcal{A}_k$ and for all $k \in \mathcal{N}$. As such, if $T = 0$, we will have $u_{k\alpha}(q) = u_{k\beta}(q)$ for all $\alpha, \beta \in \mathcal{A}_k$, i.e. q will be a Nash equilibrium of $\mathfrak{G}$; otherwise, for $T > 0$, a comparison with the KKT conditions (2.13) implies that q is the (unique) solution of the maximization problem:

$$\begin{aligned} q_k &= \arg\max_{x_k \in \mathcal{X}_k} \left\{ \sum\nolimits_\beta^k x_{k\beta} u_{k\beta}(x_k; q_{-k}) - T h_k(x_k) \right\} \\ &= \arg\max_{x_k \in \mathcal{X}_k} \left\{ \sum\nolimits_\beta^k x_{k\beta} \cdot T^{-1} u_{k\beta}(x_k; q_{-k}) - h_k(x_k) \right\} = Q_k\left(T^{-1} u_k(q)\right), \end{aligned} \tag{3.1}$$

i.e. q is a Q-equilibrium of $\mathfrak{G}$ with rationality level $\rho = 1/T$. □

Proposition 3.2 shows that the discount rate T of the dynamics (PD) plays a double role: on the one hand, it determines the discount rate of the players' assessment phase (2.2), so it reflects the importance that players give to past observations; on the other hand, T also determines the rationality level of the rest points of (PD), measuring how far the stationary points of the players' learning process are from being Nash. That being said, stationarity does not capture the long-run behavior of a dynamical system, so the rest of our analysis will be focused on the asymptotic properties of (PD). To that end, we begin with the special case of potential games where the players' payoff functions are aligned along a potential function in the sense of (1.2); in this context, the game's potential function is "almost" increasing along the solution orbits of (PD) if T is small enough:

Lemma 3.3. *Let $\mathfrak{G} \equiv \mathfrak{G}(\mathcal{N}, \mathcal{A}, u)$ be a finite game with potential U and assume that each player $k \in \mathcal{N}$ is endowed with a decomposable penalty function $h_k : \mathcal{X}_k \to \mathbb{R}$. Then, the function*

$$F(x) \equiv T \sum_{k \in \mathcal{N}} h_k(x_k) - U(x) \tag{3.2}$$

is Lyapunov for the penalty-regulated dynamics (PD): for any interior orbit $x(t)$ of (PD), we have $\frac{d}{dt} F(x(t)) \leq 0$ with equality if and only if $x(0)$ is a QRE of $\mathfrak{G}$.

Proof. Proof. By differentiating F, we readily obtain:

$$\frac{\partial F}{\partial x_{k\alpha}} = T\theta_k'(x_{k\alpha}) - u_{k\alpha}(x), \tag{3.3}$$

where θ_k is the kernel of the penalty function of player k and we have used the potential property (1.2) of $\mathfrak{G}$ to write $\frac{\partial U}{\partial x_{k\alpha}} = u_{k\alpha}$. Hence, for any interior orbit $x(t)$ of (PD), some algebra yields:

$$\begin{aligned}\frac{dF}{dt} &= \sum\nolimits_k \sum\nolimits_\alpha^k \frac{\partial F}{\partial x_{k\alpha}} \dot{x}_{k\alpha} \\ &= -\sum\nolimits_k \sum\nolimits_\alpha^k \frac{1}{\theta''(x_{k\alpha})} \left(T\theta_k'(x_{k\alpha}) - u_{k\alpha}(x)\right)^2 \\ &\quad + \sum\nolimits_k \Theta_k''(x_k) \left[\sum\nolimits_\alpha^k \frac{1}{\theta_k''(x_{k\alpha})} \left(T\theta_k'(x_{k\alpha}) - u_{k\alpha}(x)\right)\right]^2 \\ &= -\sum\nolimits_k \frac{1}{\Theta_k''(x_k)} \left[\sum\nolimits_\alpha^k \pi_{k\alpha} w_{k\alpha}^2 - \left(\sum\nolimits_\alpha^k \pi_{k\alpha} w_{k\alpha}\right)^2\right], \end{aligned} \tag{3.4}$$

where we have set $\pi_{k\alpha} = \Theta_k''(x_{k\alpha})/\theta_k''(x_{k\alpha})$ and $w_{k\alpha} = T\theta_k'(x_{k\alpha}) - u_{k\alpha}(x)$. Since $\pi_{k\alpha} \geq 0$ and $\sum_\alpha^k \pi_{k\alpha} = 1$ by construction, our assertion follows by Jensen's inequality (simply note that the condition $w_{k\alpha} = w_{k\beta}$ for all $\alpha, \beta \in \mathcal{A}_k$ is only satisfied at the QRE of $\mathfrak{G}$). □

Needless to say, Lemma 3.3 can be easily extended to orbits lying in any subface $\mathcal{X}'$ of $\mathcal{X}$ by considering the game's restricted QRE. Indeed, given that the restricted QRE of $\mathfrak{G}$ that are supported in a subface $\mathcal{X}'$ of $\mathcal{X}$ coincide with the local minimizers of $F|_{\mathcal{X}'}$, Lemma 3.3 gives:

Proposition 3.4. *Let $x(t)$ be a solution orbit of the penalty-regulated dynamics (PD) for a potential game $\mathfrak{G}$. Then:*

1. *For $T > 0$, $x(t)$ converges to a restricted QRE of $\mathfrak{G}$ with the same support as $x(0)$.*
2. *For $T = 0$, $x(t)$ converges to a restricted Nash equilibrium with support contained in that of $x(0)$.*

Proposition 3.4 implies that interior solutions of (PD) for $T > 0$ can only converge to interior points in potential games; as we show below, this behavior actually applies to *any* finite game:

Proposition 3.5. *Let $x(t)$ be an interior solution orbit of the penalty-regulated dynamics (PD) for $T > 0$. Then, any ω-limit of $x(t)$ is interior; in particular, the boundary $\operatorname{bd}(\mathcal{X})$ of $\mathcal{X}$ repels all interior orbits.*[9]

Proof. Proof. Our proof will be based on the integral representation (PRL) of the penalty-regulated dynamics (PD). Indeed, with $u_{k\alpha}$ bounded on $\mathcal{X}$ (say by some $M > 0$), we get:

$$\begin{aligned}|y_{k\alpha}(t)| \leq |y_{k\alpha}(0)|\, e^{-Tt} + \int_0^t e^{-T(t-s)}\, |u_{k\alpha}(x(s))|\, ds \\ \leq |y_{k\alpha}(0)|\, e^{-Tt} + \frac{M}{T}\left(1 - e^{-Tt}\right), \end{aligned} \tag{3.5}$$

so any ω-limit of (PRL) must lie in the rectangle $C^T = \prod_k C_k^T$ where $C_k^T = \{y_k \in \mathbb{R}^{\mathcal{A}_k} : |y_{k\alpha}| \leq M/T\}$. However, since Q_k maps $\mathbb{R}^{\mathcal{A}_k}$ to $\operatorname{rel\,int}(\mathcal{X}_k)$ continuously,

[9] Of course, orbits that start on $\operatorname{bd}(\mathcal{X})$ will remain in $\operatorname{bd}(\mathcal{X})$ for all $t \geq 0$.

$Q_k(C_k^T)$ will be a compact set contained in $\operatorname{rel\,int}(\mathcal{X}_k)$, and our assertion follows by recalling that $x(t) = Q(y(t))$. □

The above highlights an important connection between the score variables $y_{k\alpha}$ and the players' mixed strategy shares $x_{k\alpha}$: the asymptotic boundedness of the scores implies that the solution orbits of (PD) will be repelled by the boundary $\operatorname{bd}(\mathcal{X})$ of the game's strategy space. On the other hand, this connection is not a two-way street because the smooth best response map $Q_k\colon \mathbb{R}^{\mathcal{A}_k} \to \mathcal{X}_k$ is not a diffeomorphism: $Q_k(y) = Q_k(y + c(1,\dots,1))$ for every $c \in \mathbb{R}$, so Q_k collapses the directions that are parallel to $(1,\dots,1)$.

To obtain a diffeomorphic set of score-like variables, let $\mathcal{A}_k = \{\alpha_{k,0}, \alpha_{k,1}, \dots\}$ denote the action set of player k and consider the *relative scores*:

$$z_{k\mu} = \theta_k'(x_{k\mu}) - \theta_k'(x_{k,0}) = y_{k\mu} - y_{k,0}, \quad \mu = 1, 2, \dots, \tag{3.6}$$

where the last equality follows from the KKT conditions (2.13). In words, $z_{k\mu}$ simply measures the score difference between the μ-th action of player k and the "flagged" 0-th action; as such, the evolution of $z_{k\mu}$ over time will be:

$$\dot{z}_{k\mu} = \dot{y}_{k\mu} - \dot{y}_{k,0} = u_{k\mu} - Ty_{k\mu} - (u_{k,0} - Ty_{k,0}) = \Delta u_{k\mu} - Tz_{k\mu}, \tag{3.7}$$

where $\Delta u_{k\mu} = u_{k\mu} - u_{k,0}$. In particular, these relative scores remain unchanged if a players' payoffs are offset by the same amount, a fact which is reflected in the following:

Lemma 3.6. *Let $\mathcal{A}_{k,0} = \mathcal{A}_k \setminus \{\alpha_{k,0}\} = \{\alpha_{k,1}, \alpha_{k,2}, \dots\}$. Then, with notation as above, the map $\iota_k\colon x_k \mapsto z_k$ is a diffeomorphism from $\operatorname{rel\,int}(\mathcal{X}_k)$ to $\mathbb{R}^{\mathcal{A}_{k,0}}$.*

Proof. Proof. We begin by showing that ι_k is surjective. Indeed, let $z_k \in \mathbb{R}^{\mathcal{A}_{k,0}}$ and set $y_k = (0, z_{k,0}, z_{k,1}, \dots)$. Then, if $x_k = Q_k(y_k)$, the KKT conditions (2.13) become $-\theta_k'(x_{k,0}) = \zeta_k$ and $z_{k\mu} - \theta_k'(x_{k\mu}) = \zeta_k$ for all $\mu \in \mathcal{A}_{k,0}$. This gives $z_{k\mu} = \theta_k'(x_{k\mu}) - \theta_k'(x_{k,0})$ for all $\mu \in \mathcal{A}_{k,0}$, i.e. ι_k is onto.

Assume now that $\theta_k'(x_{k\mu}) - \theta_k'(x_{k\mu}) = \theta_k'(x'_{k\mu}) - \theta_k'(x'_{k,0})$ for some $x_k, x'_k \in \operatorname{rel\,int}(\mathcal{X}_k)$. A trivial rearrangement gives $\theta_k'(x_{k\alpha}) - \theta_k'(x'_{k\alpha}) = \theta_k'(x_{k\beta}) - \theta_k'(x'_{k\beta})$ for all $\alpha, \beta \in \mathcal{A}_k$, so there exists some $\xi_k \in \mathbb{R}$ such that $\theta_k'(x'_{k\alpha}) = \xi_k + \theta_k'(x_{k\alpha})$ for all $\alpha \in \mathcal{A}_k$. With θ_k' strictly increasing, this implies that $x'_{k\alpha} > x_{k\alpha}$ (resp. $x'_{k\alpha} < x_{k\alpha}$, resp. $x'_{k\alpha} = x_{k\alpha}$) for all $\alpha \in \mathcal{A}_k$ if $\xi_k > 0$ (resp. $\xi_k < 0$, resp. $\xi_k = 0$). However, given that the components of x_k and x'_k both sum to 1, we must have $x'_{k\alpha} = x_{k\alpha}$ for all α i.e. the map $x_k \mapsto z_k$ is injective.

Now, treating $x_{k,0} = 1 - \sum_{\mu\in\mathcal{A}_{k,0}} x_{k\mu}$ as a dependent variable, the Jacobian matrix of ι_k will be:

$$J^k_{\mu\nu} = \frac{\partial z_{k\mu}}{\partial x_{k\nu}} = \theta_k''(x_{k\mu})\delta_{\mu\nu} + \theta_k'\left(1 - \textstyle\sum_{\mu\in\mathcal{A}_{k,0}} x_{k\mu}\right). \tag{3.8}$$

Then, letting $\theta''_{k\mu} = \theta_k''(x_{k\mu})$ and $\theta''_{k,0} = \theta_k'\left(1 - \sum_\mu^k x_{k\mu}\right)$, it is easy to see that $J^k_{\mu\nu}$ is invertible with inverse matrix

$$J_k^{\mu\nu} = \frac{\delta_{\mu\nu}}{\theta''_{k\mu}} - \frac{\Theta_k''}{\theta''_{k\mu}\theta''_{k\nu}}, \tag{3.9}$$

where $\Theta''_k = \big(\sum_{\alpha\in\mathcal{A}_k} 1/\theta''_{k\alpha}\big)^{-1}$. Indeed, dropping the index k for simplicity, a simple inspection gives:

$$\begin{aligned}\sum_{\nu\neq 0} J_{\mu\nu}J^{\nu\rho} &= \sum_{\nu\neq 0} \big(\theta''_\mu\delta_{\mu\nu} + \theta''_0\big)\cdot\big(\delta_{\nu\rho}/\theta''_\nu - \Theta''/(\theta''_\nu\theta''_\rho)\big)\\ &= \sum_{\nu\neq 0} \big(\theta''_\mu\delta_{\mu\nu}\delta_{\nu\rho}/\theta''_\nu + \theta''_0\delta_{\nu\rho}/\theta''_\nu - \theta''_\mu\Theta''\delta_{\mu\nu}/(\theta''_\nu\theta''_\rho) - \theta''_0\Theta''/(\theta''_\nu\theta''_\rho)\big)\\ &= \delta_{\mu\rho} + \theta''_0/\theta''_\rho - \Theta''/\theta''_\rho - \theta''_0\Theta''\sum_{\nu\neq 0} 1/(\theta''_\nu\theta''_\rho) = \delta_{\mu\rho}. \end{aligned}\tag{3.10}$$

The above shows that ι_k is a smooth immersion; since ι_k is bijective, it will also be a diffeomorphism by the inverse function theorem, and our proof is complete. □

With this diffeomorphism at hand, we now show that the penalty-regulated dynamics (PD) are contracting if $T > 0$ (a result which ties in well with Proposition 3.5 above):

Proposition 3.7. *Let $K_0 \subseteq \operatorname{rel\,int}(\mathcal{X})$ be a compact set of interior initial conditions and let $K_t = \{x(t) : x(0) \in K_0\}$ be its evolution under the dynamics (PD). Then, there exists a volume form* Vol *on* $\operatorname{rel\,int}(\mathcal{X})$ *such that*

$$\operatorname{Vol}(K_t) = \operatorname{Vol}(K_0)\exp(-TA_0t),\tag{3.11}$$

where $A_0 = \sum_k(\operatorname{card}(\mathcal{A}_k) - 1)$. In other words, the penalty-regulated dynamics (PD) are incompressible for $T = 0$ and contracting for $T > 0$.

Proof. Proof. Our proof will be based on the relative score variables $z_{k\mu}$ of (3.6). Indeed, let U_0 be an open set of $\prod_k \mathbb{R}^{\mathcal{A}_{k,0}}$ and let $W_{k\mu} = \Delta u_{k\mu}(x) - Tz_{k\mu}$ denote the RHS of (3.7). Liouville's theorem then gives

$$\frac{d}{dt}\operatorname{Vol}_0(U_t) = \int_{U_t} \operatorname{div} W\,\mathrm{d}\Omega_0,\tag{3.12}$$

where $\mathrm{d}\Omega_0 = \bigwedge_{k,\mu} dz_{k\mu}$ is the ordinary Euclidean volume form on $\prod_k \mathbb{R}^{\mathcal{A}_{k,0}}$, Vol_0 denotes the associated (Lebesgue) measure on $\prod_k \mathbb{R}^{\mathcal{A}_{k,0}}$ and U_t is the image of U_0 at time t under (3.7). However, given that $\Delta u_{k\mu}$ does not depend on z_k (recall that $u_{k\mu}$ and $u_{k,0}$ themselves do not depend on x_k), we will also have $\frac{\partial W_{k\mu}}{\partial z_{k\mu}} = -T$. Hence, summing over all $\mu \in \mathcal{A}_{k,0}$ and $k \in \mathcal{N}$, we obtain $\operatorname{div} W = -\sum_k(\operatorname{card}(\mathcal{A}_k) - 1)T = -A_0T$ and (3.12) yields $\operatorname{Vol}(U_t) = \operatorname{Vol}(U_0)\exp(-A_0Tt)$.

In view of the above, let $\iota = (\iota_1, \dots, \iota_N)\colon \operatorname{rel\,int}(\mathcal{X}) \to \prod_k \mathbb{R}^{\mathcal{A}_{k,0}}$ be the product of the "relative score" diffeomorphisms of Lemma 3.6, and let $\operatorname{Vol} = \iota^*\operatorname{Vol}_0$ be the pullback of the Euclidean volume $\operatorname{Vol}_0(\cdot)$ on $\prod_k \mathbb{R}^{\mathcal{A}_{k,0}}$ to $\operatorname{rel\,int}(\mathcal{X})$, i.e. $\operatorname{Vol}(K) = \operatorname{Vol}_0(\iota(K))$ for any (Borel) $K \subseteq \operatorname{rel\,int}(\mathcal{X})$. Then, letting $U_0 = \iota(K_0)$, our assertion follows from the volume evolution equation above and the fact that $\iota(x(t))$ solves (3.7) whenever $x(t)$ solves (PD). □

When applied to (RD_T) for $T = 0$, Proposition 3.7 yields the classical result that the asymmetric replicator dynamics (RD) are incompressible – and thus do not admit interior attractors (Hofbauer and Sigmund [15], Ritzberger and Weibull [31]).[10] We thus see that incompressibility characterizes a much more general class

[10]This does not hold in the symmetric case because the symmetrized payoff $u_\alpha(x)$ depends on x_α.

of dynamics: in our learning context, it simply reflects the fact that players weigh their past observations uniformly (neither discounting, nor reinforcing them).

That said, in the case of the replicator dynamics, we have a significantly clearer picture regarding the stability and attraction properties of a game's equilibria; in particular, the *folk theorem of evolutionary game theory* (Hofbauer and Sigmund [15]) states that:[11]

(1) If an interior trajectory converges, its limit is Nash.
(2) If a state is Lyapunov stable, then it is also Nash.
(3) A state is asymptotically stable if and only if it is a strict Nash equilibrium.

By comparison, in the context of the penalty-regulated game dynamics (PD), we have:

Theorem 3.8. *Let $\mathfrak{G} \equiv \mathfrak{G}(\mathcal{N}, \mathcal{A}, u)$ be a finite game, let $h_k \colon \mathcal{X}_k \to \mathbb{R}$ be a decomposable penalty function for each player $k \in \mathcal{N}$, and let $Q_k \colon \mathbb{R}^{\mathcal{A}_k} \to \mathcal{X}_k$ denote each player's choice map. Then, the penalty-regulated dynamics (PD) have the following properties:*

(1) For $T > 0$, if $q \in \mathcal{X}$ is Lyapunov stable then it is also a QRE of $\mathfrak{G}$; moreover, if q is a Q-approximate strict Nash equilibrium and T is small enough, then q is also asymptotically stable.
(2) For $T = 0$, if $q \in \mathcal{X}$ is Lyapunov stable, then it is also a Nash equilibrium of $\mathfrak{G}$; furthermore, q is asymptotically stable if and only if it is a strict Nash equilibrium of $\mathfrak{G}$.

Proof. Proof. Our proof will be broken up in two parts depending on the discount rate T of (PD):
The case $T > 0$. Let $T > 0$ and assume that $q \in \mathcal{X}$ is Lyapunov stable (and, hence, stationary). Clearly, if q is interior, it must also be a QRE of $\mathfrak{G}$ by Proposition 3.2, so there is nothing to show. Suppose therefore that $q \in \operatorname{bd}(\mathcal{X})$; then, by Proposition 3.5, we may pick a neighborhood U of q in $\mathcal{X}$ such that $\operatorname{cl}(U)$ does not contain any ω-limit points of the interior of $\mathcal{X}$ under (PD). However, since q is Lyapunov stable, any interior solution that is wholly contained in U must have an ω-limit in $\operatorname{cl}(U)$, a contradiction.

Regarding the asymptotic stability of Q-approximate strict equilibria, assume without loss of generality that $q^* = (\alpha_{1,0}, \dots, \alpha_{N,0})$ is a strict Nash equilibrium of $\mathfrak{G}$ and let $q \equiv q(T) \in \mathcal{X}$ be a Q-approximation of q^* with rationality level $\rho = 1/T$. Furthermore, let $W_{k\mu} = \Delta u_{k\mu} - T z_{k\mu}$ and consider $\Delta u_{k\mu}$ as a function of only $x_{\ell\mu}$, $\mu \in \mathcal{A}_{\ell,0}$, by treating $x_{\ell,0} = 1 - \sum_\mu x_{\ell\mu}$ as a dependent variable. Then, as in the proof of Lemma 3.6, a simple differentiation yields:

$$\left.\frac{\partial W_{k\mu}}{\partial z_{\ell\nu}}\right|_q = \begin{cases} -T & \text{if } \ell = k,\ \nu = \mu, \\ 0 & \text{if } \ell = k,\ \nu \neq \mu, \\ \sum_{\rho \in \mathcal{A}_{\ell,0}} J_\ell^{\nu\rho}(q) \frac{\partial}{\partial w_{\ell\rho}} \Delta u_{k\mu} & \text{otherwise,} \end{cases} \tag{3.13}$$

where $J_\ell^{\nu\rho}(q)$ denotes the inverse Jacobian matrix (3.9) of the map $x \mapsto z$ evaluated at q.

[11] Recall that $q \in \mathcal{X}$ is said to be *Lyapunov stable* (or *stable*) when for every neighborhood U of q in $\mathcal{X}$, there exists a neighborhood V of q in $\mathcal{X}$ such that if $x(0) \in V$ then $x(t) \in U$ for all $t \geq 0$; q is called *attracting* when there exists a neighborhood U of q in $\mathcal{X}$ such that $\lim_{t\to\infty} x(t) = q$ if $x(0) \in U$; finally, q is called *asymptotically stable* when it is both stable and attracting.

We will show that all the elements of (3.13) with $\ell \neq k$ or $\mu \neq \nu$ are of order $o(T)$ as $T \to 0^+$, so (3.13) is dominated by the diagonal elements $\frac{\partial W_{k\mu}}{\partial z_{k\mu}} = -T$ for small T. To do so, it suffices to show that $T^{-1} J_\ell^{\nu\rho} \to +\infty$ as $T \to 0^+$; however, since q is a Q-approximation of the strict equilibrium $q^* = (\alpha_{1,0}, \dots, \alpha_{N,0})$, we will also have $q_{k\mu} \equiv q_{k\mu}(T) \to q^*_{k\mu} = 0$ and $q_{k,0} \to q^*_{k,0} = 1$ as $T \to 0^+$. Moreover, recalling that q is a QRE of $\mathfrak{G}$ with rationality level $\rho = 1/T$, we will also have $\Delta u_{k\mu}(q) = T\theta'(q_{k\mu}) - T\theta'(q_{k,0})$, implying in turn that $T\theta'(q_{k\mu}(T)) \to \Delta u_{k\mu}(q^*) < 0$ as $T \to 0^+$. We thus obtain:

$$\frac{1}{T\theta''(q_{k\mu}(T))} = \frac{\theta'(q_{k\mu}(T))}{\theta''(q_{k\mu}(T))} \frac{1}{T\theta'(q_{k\mu}(T))} \to \frac{0}{\Delta u_{k\mu}(q^*)} = 0, \tag{3.14}$$

and hence, on account of (3.9) and (3.14), we will have $J_\ell^{\nu\rho} = o(T)$ for small T. By continuity, the eigenvalues of (3.13) evaluated at $q \equiv q(T)$ will all be negative if $T > 0$ is small enough, so q will be a hyperbolic rest point of (3.7); by the Hartman-Grobman theorem it will then also be structurally stable, and hence asymptotically stable as well.

The case $T = 0$. For $T = 0$, let q be Lyapunov stable so that every neighborhood U of q in $\mathcal{X}$ admits an interior orbit $x(t)$ that stays in U for all $t \geq 0$; we then claim that q is Nash. Indeed, assume ad abusrdum that $\alpha_{k,0} \in \operatorname{supp}(q)$ has $u_{k,0}(q) < u_{k\mu}(q)$ for some $\mu \in \mathcal{A}_{k,0} \equiv \mathcal{A}_k \setminus \{\alpha_{k,0}\}$, and let U be a neighborhood of q such that $x_{k,0} > q_{k,0}/2$ and $\Delta u_{k\mu}(x) \geq m > 0$ for all $x \in U$. Picking an orbit $x(t)$ that is wholly contained in U, the dynamics (3.7) readily give $z_{k\mu}(t) \geq z_{k,0}(0) + mt$, implying in turn that $z_{k\mu}(t) \to +\infty$ as $t \to \infty$. However, with $z_{k\mu} = \theta'(x_{k\mu}) - \theta'(x_{k,0})$, this is only possible if $x_{k\mu}(t) \to 0$, a contradiction.

Assume now that $q = (\alpha_{1,0}, \dots, \alpha_{N,0})$ is a strict Nash equilibrium of $\mathfrak{G}$. To show that q is Lyapunov stable, it will be again convenient to work with the relative scores $z_{k\mu}$ and show that if $m \in \mathbb{R}$ is sufficiently negative, then every trajectory $z(t)$ that starts in the open set $U_m = \{z \in \prod_k \mathbb{R}^{\mathcal{A}_{k,0}} : z_{k\mu} < m\}$ always stays in U_m; since U_m maps via $\iota^{-1} \colon \prod_k \mathbb{R}^{\mathcal{A}_{k,0}} \to \operatorname{rel\,int}(\mathcal{X})$ to a neighborhood of q in $\operatorname{rel\,int}(\mathcal{X})$, this is easily seen to imply Lyapunov stability for q in $\mathcal{X}$.

In view of the above, pick $m \in \mathbb{R}$ so that $\Delta u_{k\mu}(x(z)) \leq -\varepsilon < 0$ for all $z \in U_m$ and let $\tau_m = \inf\{t : z(t) \notin U_m\}$ be the time it takes $z(t)$ to escape U_m. Then, if τ_m is finite and $t \leq \tau_m$, the relative score dynamics (3.7) readily yield

$$z_{k\mu}(t) = z_{k\mu}(0) + \int_0^t \Delta u_{k\mu}(Q_0(z(s)))ds \leq z_{k\mu}(0) - \varepsilon t < m \quad \text{for all } \mu \in \mathcal{A}_{k,0},\ k \in \mathcal{N}. \tag{3.15}$$

Thus, substituting τ_m for t in (3.15), we obtain a contradiction to the definition of τ_m and we conclude that $z(t)$ always stays in U_m if m is chosen negative enough – i.e. q is Lyapunov stable.

To show that q is in addition attracting, it suffices to let $t \to \infty$ in (3.15) and recall the definition (3.6) of the $z_{k\mu}$ variables. Finally, for the converse implication, assume that q is not pure; in particular, assume that q lies in the relative interior of a non-singleton subface $\mathcal{X}'$ spanned by $\operatorname{supp}(q)$. Proposition 3.7 shows that q cannot attract a relatively open neighborhood U' of initial conditions in $\mathcal{X}'$ because (PD) remains volume-preserving when restricted to any subface $\mathcal{X}'$ of $\mathcal{X}$. This implies that q cannot be attracting in $\mathcal{X}$, so q cannot be asymptotically stable either. □

In conjunction with our previous results, Theorem 3.8 provides an interesting insight into the role of the dynamics' discount rate T: for small $T > 0$, the dynamics (PD) are attracted to the interior of $\mathcal{X}$ and can only converge to points that are *approximately* Nash; on the other hand, for $T = 0$, the solutions (PD) are only attracted to strict Nash equilibria (see also Fig. 1). As such, Theorem 3.8 and Proposition 3.4 suggest that if one seeks to reach a (pure) Nash equilibrium, the best convergence properties are provided by the "no discounting" case $T = 0$. Nonetheless, as we shall see in the following section, if one seeks to implement the dynamics (PD) as a bona fide learning algorithm in discrete time, the "positive discounting" regime $T > 0$ is much more robust than the "no disounting" case – all the while allowing players to converge arbitrarily close to a Nash equilibrium.

3.2. **The case $T < 0$: reinforcing past observations.** In this section, we examine briefly what happens when players use a negative discount rate $T < 0$, i.e. they reinforce past observations instead of discounting them. As we shall see, even though the form of the dynamics (PRL)/(PD) remains the same (the derivation of (PRL) and (PD) does not depend on the sign of T), their properties are quite different in the regime $T < 0$.

The first thing to note is that the definition of a QRE also extends to negative rationality levels $\rho < 0$ that describe an "anti-rational" behavior where players attempt to minimize their payoffs: indeed, the QRE of a game $\mathfrak{G} \equiv \mathfrak{G}(\mathcal{N}, \mathcal{A}, u)$ for negative ρ are simply QRE of the opposite game $-\mathfrak{G} \equiv (\mathcal{N}, \mathcal{A}, -u)$, and as $\rho \to -\infty$, these equilibria approximate the Nash equilibria of $-\mathfrak{G}$.

In this way, repeating the analysis of Section 3.1, we obtain:

Theorem 3.9. *Let $\mathfrak{G} \equiv \mathfrak{G}(\mathcal{N}, \mathcal{A}, u)$ be a finite game and assume that each player $k \in \mathcal{N}$ is endowed with a decomposable penalty function $h_k \colon \mathcal{X}_k \to \mathbb{R}$ with induced choice map $Q_k \colon \mathbb{R}^{\mathcal{A}_k} \to \mathcal{X}_k$. Then, in the case of a negative discount rate $T < 0$:*

(*1*) *The rest points of the penalty-regulated dynamics* (PD) *are the restricted QRE of the opposite game $-\mathfrak{G}$.*
(*2*) *The dynamics* (PD) *are expanding with respect to the volume form of Proposition 3.7 and* (3.11) *continues to hold.*
(*3*) *A strategy profile $q \in \mathcal{X}$ is asymptotically stable if and only if it is pure (i.e. a vertex of $\mathcal{X}$); any other rest point of* (PD) *is unstable.*

Theorem 3.9 will be our main result for $T < 0$, so some remarks in order:

Remark 7. The games $\mathfrak{G}$ and $-\mathfrak{G}$ have the same restricted equilibria, so the rest points of (PD) for small $T > 0$ (corresponding to QRE with large $\rho = 1/T \to +\infty$) transition smoothly to perturbed equilibria with small $T < 0$ ($\rho \to -\infty$) via the "fully rational" case $T = 0$ (which corresponds to the Nash equilibria of the game when $\rho = \pm\infty$). In fact, by continuity, the phase portrait of the dynamics (PD) for sufficiently small T (positive or negative) will be broadly similar to the base case $T = 0$ (at least, in the generic case where there are no payoff ties in $\mathfrak{G}$). The main difference between positive and negative discount rates is that, for small $T < 0$, the orbits of (PD) are attracted to the vertices of $\mathcal{X}$ (though each individual vertex might have a vanishingly small basin of attraction), whereas for small $T > 0$, the dynamics are only attracted to interior points (which, however, are arbitrarily close to the vertices of $\mathcal{X}$).

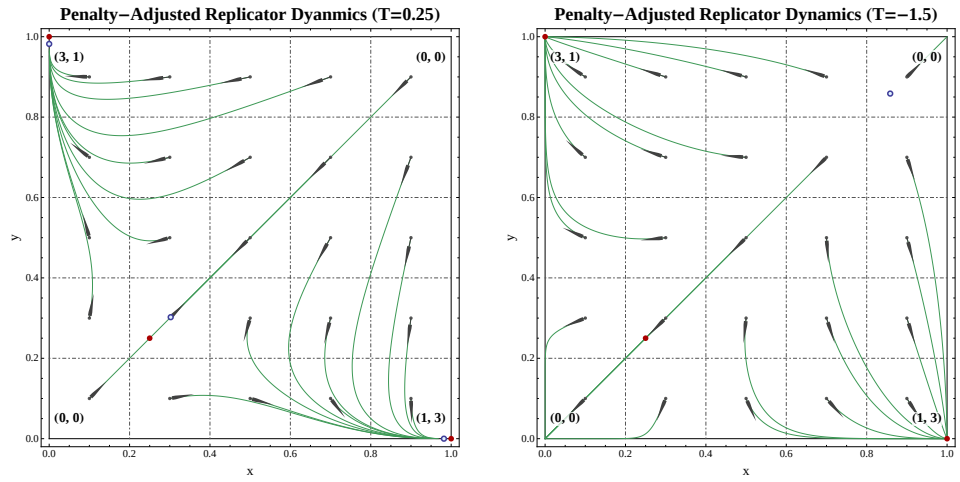


FIGURE 2. Phase portraits of the penalty-adjusted replicator dynamics (RD_T) showing the transition from positive to negative discount rates in the game of Fig. 1. For small $T > 0$, the rest points of (PD) are Q-approximate Nash equilibria (red dots), and they attract almost all interior solutions; as T drops to negative values, the non-equilibrium vertices of $\mathcal{X}$ become asymptotically stable (but with a small basin of attraction), and each one gives birth to an unstable QRE of the opposite game in a subcritical pitchfork bifurcation. Of these two equilibria, the one closer to the game's interior Nash equilibrium is annihilated with the pre-existing QRE at $T \approx -0.278$, and as $T \to -\infty$, we obtain a time-inverted image of the $T \to +\infty$ portrait with the only remaining QRE repelling all trajectories towards the vertices of $\mathcal{X}$; Figure 2(b) shows the case where only one (repelling) rest point remains.

Remark 8. It should be noted that the expanding property of (PD) for $T < 0$ does not clash with the fact that the vertices of $\mathcal{X}$ are asymptotically stable. Indeed, as can be easily seen by Lemma 3.6, sets of unit volume become vanishingly small (in the Euclidean sense) near the boundary $\mathrm{bd}(\mathcal{X})$ of $\mathcal{X}$; as such, the expanding property of the dynamics (PD) precludes the existence of attractors in the interior of $\mathcal{X}$, but not of boundary attractors.[12]

Proof. Proof of Theorem 3.9. The time inversion $t \mapsto -t$ in (PD) is equivalent to the inversion $u \mapsto -u$, $T \mapsto -T$, so our first claim follows from the $T > 0$ part of Proposition 3.2; likewise, our second claim is obtained by noting that the proof of Proposition 3.7 does not differentiate between positive and negative temperatures either.

For the last part, our proof will be based on the dynamics (3.7); more precisely, focus for convenience on the vertex $q = (\alpha_{1,0}, \dots, \alpha_{N,0})$ of $\mathcal{X}$, and let $\mathcal{A}_{k,0} = \mathcal{A}_k \setminus \{\alpha_{k,0}\}$ as usual. Then, a simple integration of (3.7) yields

$$z_{k\mu}(t) = z_{k\mu}(0)e^{-Tt} + \int_0^t e^{-T(t-s)} \Delta u_{k\mu}(x(s))\, ds. \tag{3.16}$$

[12]Obviously, the same applies to every subface $\mathcal{X}'$ of $\mathcal{X}$, explaining in this way why only the vertices of $\mathcal{X}$ are attracting.

However, given that $\Delta u_{k\mu}$ is bounded on $\mathcal{X}$ (say by some $M > 0$), the last integral will be bounded in absolute value by $M\,|T|^{-1}\left(e^{|T|t}-1\right)$, and hence:

$$z_{k\mu}(t) \leq -M\,|T|^{-1} + \left(z_{k\mu}(0) + M\,|T|^{-1}\right) e^{|T|t}. \tag{3.17}$$

Thus, if we pick $z_{k\mu}(0) < -M\,|T|^{-1}$, we will have $\lim_{t\to\infty} z_{k\mu}(t) = -\infty$ for all $\mu \in \mathcal{A}_{k,0}$, $k \in \mathcal{N}$, i.e. $x(t) \to q$. Accordingly, given that the set $U_T = \{z \in \prod_k \mathbb{R}^{\mathcal{A}_{k,0}} : z_{k\mu} < -M\,|T|^{-1}\}$ is just the image of a neighborhood of q in $\operatorname{rel\,int}(\mathcal{X})$ under the diffeomorphism of Lemma 3.6, q will attract all nearby interior solutions of (PD); by restriction, this property applies to any subface of $\mathcal{X}$ which contains q, so q is attracting. Finally, if $z_{k\mu}(0) < -M\,|T|^{-1}$, we will also have $z_{k\mu}(t) < z_{k\mu}(0)$ for all $t \geq 0$ (cf. the proof of Proposition 3.5), so q is Lyapunov stable, and hence asymptotically stable as well.

Conversely, assume that $q \in \mathcal{X}$ is a non-pure Lyapunov stable state; then, by descending to a subface of $\mathcal{X}$ if necessary, we may assume that q is interior. In that case, if U is a neighborhood of q in $\operatorname{rel\,int}(\mathcal{X})$, Proposition 3.7 shows that any neighborhood V of q that is contained in U will eventually grow to a volume larger than that of U under (PD), so there is no open set of trajectories contained in U. This shows that only vertices of $\mathcal{X}$ can be stable, and our proof is complete. □

4. Discrete-time learning algorithms

In this section, we examine how the dynamics (PRL) and (PD) may be used for learning in finite games that are played repeatedly over time. To that end, a first-order Euler discretization of the dynamics (PRL) gives the recurrence

$$\begin{aligned} Y_{k\alpha}(n+1) &= Y_{k\alpha}(n) + \gamma\left[u_{k\alpha}(X(n)) - TY_{k\alpha}(n)\right], \\ X_k(n+1) &= Q(Y_k(n+1)), \end{aligned} \tag{4.1}$$

which is well-known to track (PRL) arbitrarily well over finite time horizons when the discretization step γ is sufficiently small. That said, in many practical scenarios, players cannot monitor the mixed strategies of their opponents, so (4.1) cannot be updated directly. As a result, in the absence of perfect monitoring (or a similar oracle-like device), any distributed discretization of the dynamics (PRL)/(PD) should involve only the players' observed payoffs and no other information.

In what follows (cf. Table 1 for a summary), we will drop such information and coordination assumptions one by one: in Algorithm 1, players will only be assumed to possess a bounded, unbiased estimate of their actions' payoffs; this assumption is then dropped in Algorithm 2 which only requires players to observe their in-game payoffs (or a perturbed version thereof); finally, Algorithm 3 provides a decentralized variant of Algorithm 2 where players are no longer assumed to update their strategies in a synchronous way.

4.1. Stochastic approximation of continuous dynamics.

We begin by recalling a few general elements from the theory of stochastic approximation. Following Benaïm [3] and Borkar [6], let $\mathcal{S}$ be a finite set, and let $Z(n)$, $n \in \mathbb{N}$, be a stochastic process in $\mathbb{R}^{\mathcal{S}}$ such that

$$Z(n+1) = Z(n) + \gamma_{n+1}U(n+1), \tag{4.2}$$

where γ_n is a sequence of step sizes and $U(n)$ is a stochastic process in $\mathbb{R}^{\mathcal{S}}$ adapted to the filtration $\mathcal{F}$ of Z. Then, given a (Lipschitz) continuous vector field $f\colon \mathbb{R}^{\mathcal{S}} \to \mathbb{R}^{\mathcal{S}}$,

	Input	Uncertainties	Asynchronicities
Algorithm 1	payoff vector	✓	no
Algorithm 2	in-game payoffs	✓	no
Algorithm 3	in-game payoffs	✓	✓

Table 1. Summary of the information and coordination requirements of the learning algorithms of Section 4.

we will say that (4.2) is a *stochastic approximation* of the dynamical system

$$\dot{z} = f(z), \tag{MD}$$

if $\mathbb{E}[U(n+1) \mid \mathcal{F}_n] = f(Z(n))$ for all n. More explicitly, if we split the so-called *innovation term* $U(n)$ of (4.2) into its average value $f(Z(n)) = \mathbb{E}[U(n+1)|\mathcal{F}_n]$ and a zero-mean noise term $V(n+1) = U(n+1) - f(Z(n))$, (4.2) takes the form

$$Z(n+1) = Z(n) + \gamma_{n+1}\left[f(Z(n)) + V(n+1)\right], \tag{SA}$$

which is just a noisy Euler-like discretization of (MD); conversely, (MD) will be referred to as the *mean dynamics* of the stochastic recursion (SA).

The main goal of the theory of stochastic approximation is to relate the process (SA) to the solution trajectories of the mean dynamics (MD). Some standard assumptions that enable this comparison are:

(A1) The step sequence γ_n is $(\ell^2 - \ell^1)$–summable, viz. $\sum_n \gamma_n = \infty$ and $\sum_n \gamma_n^2 < \infty$.

(A2) $V(n)$ is a martingale difference with $\sup_n \mathbb{E}\left[\|V(n+1)\|^2 \;\middle|\; \mathcal{F}_n\right] < \infty$.

(A3) The stochastic process $Z(n)$ is bounded: $\sup_n \|Z(n)\| < \infty$ (a.s.).

Under these assumptions, the next lemma provides a sufficient condition which ensures that (SA) converges to the set of stationary points of (MD):

Lemma 4.1. *Assume that the dynamics* (MD) *admit a strict Lyapunov function* (*i.e. a real-valued function which decreases along non-stationary orbits of* (MD)) *such that the set of values taken by this function at the rest points of* (MD) *has measure zero in* $\mathbb{R}$. *Then, under Assumptions* (A1)–(A3) *above, every limit point of the stochastic approximation process* (SA) *belongs to a connected set of rest points of the mean dynamics* (MD).

Proof. Proof. Our claim is a direct consequence of the following string of results in Benaïm [3]: Prop. 4.2, Prop. 4.1, Theorem 5.7, and Prop. 6.4. □

As an immediate application of Lemma 4.1, let $\mathfrak{G}$ be a finite game with potential U. By Lemma 3.3, the function $F = Th - U$ is Lyapunov for (PRL)/(PD); moreover, since U is multilinear and h is smooth and strictly convex, Sard's theorem (Lee [22]) ensures that the set of values taken by F at its critical points has measure zero. Thus, any stochastic approximation of (PRL)/(PD) which satisfies Assumptions (A1)–(A3) above can only converge to a connected set of restricted QRE.

4.2. **Score-based learning.** In this section, we present an algorithmic implementation of the score-based learning dynamics (PRL) under two different information assumptions: first, we will assume that players possess an unbiased estimate for the payoff of each of their actions (including those that they did not play at a given instance); we will then drop this assumption and describe the issues that arise when players can only observe their in-game payoffs.

4.2.1. *Learning with imperfect payoff estimates.* If the players can estimate the payoffs of actions that they did not play, the sequence of play will be as follows:

(1) At stage $n+1$, each player selects an action $\alpha_k(n+1) \in \mathcal{A}_k$ based on a mixed strategy $X_k(n) \in \mathcal{X}_k$.
(2) Every player receives a bounded and unbiased estimate $\hat{u}_{k\alpha}(n+1)$ of his actions' payoffs, viz.
   (a) $\mathbb{E}\left[\hat{u}_{k\alpha}(n+1) \mid \mathcal{F}_n\right] = u_{k\alpha}(X(n))$,
   (b) $|\hat{u}_{k\alpha}(n+1)| \leq C$ (a.s.),
   where $\mathcal{F}_n$ is the history of the process up to stage n and $C > 0$ is a constant.
(3) Players choose a mixed strategy $X_k(n+1) \in \mathcal{X}_k$ and the process repeats.

It should be noted here that players are not explicitly assumed to monitor their opponents' strategies, nor to communicate with each other in any way: for instance, in congestion and resource allocation games, players can compute their out-of-game payoffs by probing the game's facilities for a broadcast. That said, the specifics of how such estimates can be obtained will not concern us here: in what follows, we only seek to examine how players can exploit such information when it is available. To that end, the score-based learning process (PRL) gives:

Algorithm 1 Score-based learning with imperfect payoff monitoring

$n \leftarrow 0$;
foreach player $k \in \mathcal{N}$ **do**
  initialize $Y_k \in \mathbb{R}^{\mathcal{A}_k}$ and set $X_k \leftarrow Q_k(Y_k)$; # initialization
Repeat
  $n \leftarrow n+1$;
  foreach player $k \in \mathcal{N}$ **do** simultaneously
  select new action $\alpha_k \in \mathcal{A}_k$ according to mixed strategy X_k; # choose action
  foreach player $k \in \mathcal{N}$ **do**
    foreach action $\alpha \in \mathcal{A}_k$ **do**
      observe $\hat{u}_{k\alpha}$; # estimate payoff of each action
      $Y_{k\alpha} \leftarrow Y_{k\alpha} + \gamma_n(\hat{u}_{k\alpha} - TY_{k\alpha})$; # update score of each action
    $X_k \leftarrow Q_k(Y_k)$; # update mixed strategy
  until termination criterion is reached

To study the convergence properties of Algorithm 1, let $Y_k(n)$ denote the score vector of player k at the n-th iteration of the algorithm – and likewise for the player's mixed strategy $X_k(n) \in \mathcal{X}_k$, chosen action $\alpha_k(n) \in \mathcal{A}_k$ and payoff estimates

$\hat{u}_{k\alpha}(n) \in \mathbb{R}$. Then, for all $k \in \mathcal{N}$ and $\alpha \in \mathcal{A}_k$, we get:

$$\begin{aligned}\mathbb{E}\left[(Y_{k\alpha}(n+1) - Y_{k\alpha}(n))/\gamma_{n+1} \mid \mathcal{F}_n\right] &= \mathbb{E}\left[\hat{u}_{k\alpha}(n+1) \mid \mathcal{F}_n\right] - TY_{k\alpha}(n) \\ &= u_{k\alpha}(X(n)) - TY_{k\alpha}(n). \end{aligned} \tag{4.3}$$

Together with the choice rule $X_k(n) = Q_k(Y_k(n))$, the RHS of (4.3) yields the score dynamics (PRL), so the process $X(n)$ generated by Algorithm 1 is a stochastic approximation of (PRL). We thus get:

Theorem 4.2. *Let $\mathfrak{G}$ be a potential game. If the step size sequence γ_n satisfies* (A1) *and the players' payoff estimates $\hat{u}_{k\alpha}$ are bounded and unbiased, Algorithm 1 converges* (a.s.) *to a connected set of QRE of $\mathfrak{G}$ with rationality parameter $\rho = 1/T$. In particular, $X(n)$ converges within $\varepsilon(T)$ of a Nash equilibrium of $\mathfrak{G}$ and the error $\varepsilon(T)$ vanishes as $T \to 0$.*

Proof. Proof. In view of the discussion following Lemma 4.1, we will establish our claim by showing that Assumptions (A1)–(A3) are all satisfied in the case of the stochastic approximation

$$Y_{k\alpha}(n+1) = Y_{k\alpha}(n) + \gamma_{n+1}\left[\hat{u}_{k\alpha}(n+1) - TY_{k\alpha}(n)\right]. \tag{4.4}$$

Assumption (A1) is true by design, so there is nothing to show. Furthermore, expressing the noise term of (4.4) as $V_{k\alpha}(n+1) = \hat{u}_{k\alpha}(n+1) - u_{k\alpha}(X(n))$, we readily obtain $\mathbb{E}\left[V_{k\alpha}(n+1) \mid \mathcal{F}_n\right] = 0$ and $\mathbb{E}\left[V_{k\alpha}^2(n+1) \,\middle|\, \mathcal{F}_n\right] \leq 2C^2$, so Assumption (A2) also holds. Finally, with $\hat{u}_{k\alpha}$ bounded (a.s.), $Y_{k\alpha}$ will also be bounded (a.s.): indeed, note first that $0 \leq 1 - T\gamma_n \leq 1$ for all n larger than some n_0; then, using the uniform norm for convenience of notation, the iterates of (4.4) will satisfy $\|Y(n+1)\| \leq (1 - T\gamma_{n+1})\|Y(n)\| + \gamma_{n+1}C$ for all sufficiently large n. Hence:

(1) If $\|Y(n)\| \leq C/T$, we will also have $\|Y(n+1)\| \leq (1 - T\gamma_{n+1})C/T + \gamma_{n+1}C = C/T$.
(2) If $\|Y(n)\| > C/T$, we will have $\|Y(n+1)\| \leq \|Y(n)\| - T\gamma_{n+1}\|Y(n)\| + \gamma_{n+1}C \leq \|Y(n)\|$, i.e. $\|Y\|$ decreases.

The above shows that $\|Y(n)\|$ is bounded by $T^{-1}C \vee \max_{n \geq n_0}\|Y(n)\|$, so Assumption (A3) also holds. By Proposition 3.2 and the discussion following Lemma 4.1, we then conclude that $X(n)$ converges to a connected set of *restricted* QRE of $\mathfrak{G}$. However, since $Y(n)$ is bounded, $X(n)$ will be bounded away from the boundary $\operatorname{bd}(\mathcal{X})$ of $\mathcal{X}$ because the image of a compact set under Q is itself compact in $\operatorname{rel\,int}(\mathcal{X})$. As such, any limit point of $X(n)$ will be interior and our claim follows. □

4.2.2. *The issue with in-game observations.* Assume now that the only information at the players' disposal is the payoff of their chosen actions, possibly perturbed by some random noise process. Formally, if $\alpha_k(n+1)$ denotes the action of player k at the $(n+1)$-th stage of the process, we will assume that the corresponding observed payoff is of the form

$$\hat{u}_k(n+1) = u_k(\alpha_1(n+1), \dots, \alpha_N(n+1)) + \xi_k(n+1), \tag{4.5}$$

where the noise process ξ_k is a bounded, $\mathcal{F}$-adapted martingale difference (i.e. $\mathbb{E}\left[\xi_k(n+1) \mid \mathcal{F}_n\right] = 0$ and $\|\xi_k\| \leq C$ for some $C > 0$) with $\xi_k(n+1)$ independent of $\alpha_k(n+1)$.[13]

[13]These assumptions are rather mild and can be easily justifed by invoking the independence between the nature-driven perturbations to the players' payoffs and the sampling done by each player to select an action at each stage. In fact, this accounts not only for i.i.d. perturbations

In this context, players only possess information regarding the actions that they actually played. Thus, motivated by the Q-learning scheme (2.4), we will use the unbiased estimator

$$\hat{u}_{k\alpha}(n+1) = \hat{u}_k(n+1)\cdot\frac{\mathbb{1}(\alpha_k(n+1)=\alpha)}{\mathbb{P}(\alpha_k(n+1)=\alpha \mid \mathcal{F}_n)} = \mathbb{1}(\alpha_k(n+1)=\alpha)\cdot\frac{\hat{u}_k(n+1)}{X_{k\alpha}(n)} \tag{4.6}$$

which allows us to replace the inner action-sweeping loop of Algorithm 1 with the update step:

---

foreach player $k \in \mathcal{N}$ **do** simultaneously
select new action $\alpha_k \in \mathcal{A}_k$ according to mixed strategy X_k; # choose action
observe $\hat{u}_k$; # receive realized payoff
$Y_{k\alpha_k} \leftarrow Y_{k\alpha_k} + \gamma_n(\hat{u}_k - TY_{k\alpha_k})/X_{k\alpha_k}$; # update current action score
$X_k \leftarrow Q(Y_k)$; # update mixed strategy

As before, $Y(n)$ is an $\mathcal{F}$-adapted process with

$$\begin{aligned}\mathbb{E}\left[(Y_{k\alpha}(n+1) - Y_{k\alpha}(n))/\gamma_{n+1}\middle|\mathcal{F}_n\right]&\\ &= \mathbb{E}\left[\mathbb{1}(\alpha_k(n+1)=\alpha)\cdot\frac{\hat{u}_k(n+1)}{X_{k\alpha}(n)}\;\middle|\;\mathcal{F}_n\right] - TY_{k\alpha}(n)\\ &= u_{k\alpha}(X(n)) - TY_{k\alpha}(n),\end{aligned} \tag{4.7}$$

where the last line follows from the assumptions on $\hat{u}_k$ and ξ_k.

The mean dynamics of (4.7) are still given by (PRL) so the resulting algorithm boils down to the Q-learning scheme of Leslie and Collins [24]. This scheme was shown to converge to a QRE (or *Nash distribution*) in several classes of 2-player games under the assumption that $Y(n)$ remains bounded, but since the probabilities $X_{k\alpha}(n)$ can become arbitrarily small, this assumption is hard to verify – so the convergence of this variant of Alg. 1 with in-game observations cannot be guaranteed either.

One possible way of overcoming the unboundedness of $Y(n)$ would be to truncate the innovation term of (4.4) with a sequence of expanding bounds as in Sharia [36]; ultimately however, the required summability conditions amount to showing that the estimator (4.6) is itself bounded, so the original difficulty remains.[14] Thus, instead of trying to show that $Y(n)$ tracks (PRL), we will focus in what follows on the strategy-based variant (PD) – which is equivalent to (PRL) in continuous time – and implement it directly as a payoff-based learning process in discrete time.

4.3. **Strategy-based learning.** In this section, we will derive an algorithmic implementation of the penalty-regulated dynamics (PD) which only requires players to observe their in-game payoffs – or a perturbed version thereof. One advantage of using (PD) as a starting point is that it does not require a closed form expression for the choice map Q (which is hard to obtain for non-logit action selection);

(a case which has attracted significant interest in the literature by itself), but also for scenarios where the noise at stage $n+1$ depends on the entire history of play up to stage n.

[14]Note that this is also true for the weaker requirement of Borkar [6], namely that the innovation term of (4.4) is bounded in L^2 by $K(1+\|Y_n\|^2)$ for some positive $K > 0$.

another is that since the algorithm is strategy-based (and hence its update variables are bounded by default), we will not need to worry too much about satisfying conditions (A2) and (A3) as in the case of Algorithm 1.

With all this in mind, we obtain the following strategy-based algorithm:

Algorithm 2 Strategy-based learning with in-game payoff observations

Parameters: $T > 0$, θ_k, γ_n
$n \leftarrow 0$;
foreach player $k \in \mathcal{N}$ **do**
  initialize $X_k \in \operatorname{rel\,int}(\mathcal{X}_k)$ as a mixed strategy with full support; # initialization
Repeat
  $n \leftarrow n + 1$;
  foreach player $k \in \mathcal{N}$ **do** simultaneously
  select new action $\alpha_k \in \mathcal{A}_k$ according to mixed strategy X_k; # choose action
  observe $\hat{u}_k$; # receive realized payoff
  foreach action $\alpha \in \mathcal{A}_k$ **do**

$$X_{k\alpha} \leftarrow X_{k\alpha} + \frac{\gamma_n}{\theta_k''(X_{k\alpha})}\left[\frac{\hat{u}_k}{X_{k\alpha_k}}\left(\mathbb{1}(\alpha_k = \alpha) - \frac{\Theta_k''(X_k)}{\theta_k''(X_{k\alpha_k})}\right) - T g_{k\alpha}(X)\right]$$

    where $g_{k\alpha}(x) \equiv \theta_k'(x_{k\alpha}) - \Theta_k''(x_k)\sum_\beta^k \theta_k'(X_{k\beta})/\theta_k''(X_{k\beta})$; # update mixed strategy
  until termination criterion is reached

Remark 1. As a specific example, the Gibbs kernel $\theta(x) = x\log x$ leads to the update rule:

$$X_{k\alpha} \leftarrow X_{k\alpha} + \gamma_n\left[\big(\mathbb{1}(\alpha_k = \alpha) - X_{k\alpha}\big)\cdot \hat{u}_k - T X_{k\alpha}\left(\log X_{k\alpha} - \textstyle\sum_\beta^k X_{k\beta}\log X_{k\beta}\right)\right]. \tag{4.8}$$

Thus, for $T = 0$, we obtain the reinforcement learning scheme of Sastry et al. [35] based on the classical replicator equation (RD).

The strategy update step of Algorithm 2 has been designed to track the dynamics (PD); indeed, for all $k \in \mathcal{N}$ and for all $\alpha \in \mathcal{A}_k$, we will have

$$\begin{aligned}\mathbb{E}&\left[(X_{k\alpha}(n+1) - X_{k\alpha}(n))/\gamma_{n+1} \mid \mathcal{F}_n\right]\\ &= \frac{1}{\theta_k''(X_{k\alpha}(n))}\left[u_{k\alpha}(X(n))\left(1 - \frac{\Theta_k''(X_k(n))}{\theta_k''(X_{k\alpha}(n))}\right) - \sum\nolimits_{\beta\neq\alpha}^k u_{k\beta}(X(n))\frac{\Theta_k''(X_k(n))}{\theta_k''(X_{k\beta}(n))}\right]\\ &\quad - \frac{T g_{k\alpha}(X(n))}{\theta_k''(X_{k\alpha}(n))},\end{aligned} \tag{4.9}$$

which is simply the RHS of (PD) evaluated at $X(n)$. On the other hand, unlike Algorithm 1 (which evolves in $\prod_k \mathbb{R}^{\mathcal{A}_k}$), Algorithm 2 is well-defined only if the iterates $X_k(n)$ are admissible mixed strategies with full support at each update step.

To check that this is indeed the case, note first that the second term of the strategy update step of Algorithm 2 vanishes when summed over $\alpha \in \mathcal{A}_k$ so $\sum_\alpha^k X_{k\alpha}(n)$ will always be equal to 1 (recall that $X_k(0)$ is initialized as a valid probability distribution); as a result, it suffices to show that $X_{k\alpha}(n) > 0$ for all $\alpha \in \mathcal{A}_k$. Normalizing

the game's payoffs to $[0,1]$ for simplicity, the next lemma shows that the iterates of Alg. 2 for $T > 0$ remain a bounded distance away from the boundary $\mathrm{bd}(\mathcal{X})$ of $\mathcal{X}$:

Lemma 4.3. *Let θ be a penalty function (cf. Definition 2.1) with $x\theta''(x) \geq m > 0$ for all $x > 0$. Then, for normalized payoff observations $\hat{u}_k \in [0,1]$ and $T > 0$, there exists a positive constant $K > 0$ (depending only on T and θ) such that the iterates $X_{k\alpha}(n)$ of Algorithm 2 remain bounded away from 0 whenever the step sequence γ_n is bounded from above by K.*

Proof. Proof. We begin with some simple facts for θ (for simplicity, we will drop the player index $k \in \mathcal{N}$ in what follows):

- θ' is strictly increasing.
- There exists some $M > 0$ such that $|\theta'(\xi)/\theta''(\xi)| < M$ for all $\xi \in (0,1)$.
- For all $x \in \mathcal{X}$, $\sum_\beta 1/\theta''(x_\beta) \geq \max_\beta 1/\theta''(x_\beta)$, so $\Theta''(x) \leq \min_\beta \theta''(x_\beta) \leq \max\{\theta''(\xi) : \mathrm{card}(\mathcal{A})^{-1} \leq \xi \leq 1\}$. In particular, there exists some $\Theta''_{\max}$ such that $0 < \Theta''(x) \leq \Theta''_{\max}$ for all $x \in \mathcal{X}$.

Now, letting $\hat{\alpha}$ be the chosen action at step $n+1$ and writing $\hat{u}$ for the corresponding observed payoff, we will have:

$$\begin{aligned}
\frac{1}{\theta''(x_\alpha)}&\left[Tg_\alpha(x) - \frac{\hat{u}}{x_{\hat{\alpha}}}\big(\mathbb{1}(\hat{\alpha}=\alpha) - \Theta''_h(x)/\theta''(x_{\hat{\alpha}})\big)\right] \\
&\leq \frac{1}{\theta''(x_\alpha)}\left[Tg_\alpha(x) + \frac{\hat{u}\Theta''(x)}{x_{\hat{\alpha}}\theta''(x_{\hat{\alpha}})}\right] \\
&\leq \frac{1}{\theta''(x_\alpha)}\left[T\left(\theta'(x_\alpha) - \Theta''(x)\textstyle\sum_\beta^k \theta'(x_\beta)/\theta''(x_\beta)\right) + m^{-1}\Theta''(x)\right] \\
&\leq \frac{1}{\theta''(x_\alpha)}\left[T\theta'(x_\alpha) + \Theta''_{\max}\left(m^{-1} + \mathrm{card}(\mathcal{A})MT\right)\right],
\end{aligned}$$

where we used the normalization $\hat{u} \in [0,1]$ in the first two lines. We thus get

$$X_\alpha(n+1) \geq X_\alpha(n) - \gamma_{n+1}X_\alpha(n)\frac{c_1\theta'(X_\alpha(n)) + c_2}{\theta''(X_\alpha(n))}, \tag{4.10}$$

where c_1 and c_2 are positive constants.

Since θ' is strictly increasing, we will have $c_1\theta'(x) + c_2 < 0$ if and only if $x < \psi_0$ for some fixed $\psi_0 \in (0,1)$ which depends only on T and θ. As such, if $X_\alpha(n) \leq \psi_0$, (4.10) gives $X_\alpha(n+1) > X_\alpha(n)$; on the other hand, if $X_\alpha(n) \geq \psi_0$, then, the coefficient of $\gamma_{n+1}X_\alpha(n)$ in (4.10) will be bounded from above by some positive constant $c > 0$ (recall that $x\theta''(x) \geq m > 0$ for all $x > 0$ and $\lim_{x\to 0^+}\theta'(x)/\theta''(x) = 0$). Therefore, if we take $K \equiv 1/(2c)$ and $\gamma_{n+1} \leq K$ for all $n \geq 0$, we readily obtain:

$$X_\alpha(n+1) \geq X_\alpha(n) - c\gamma_{n+1}X_\alpha(n) \geq \tfrac{1}{2}X_\alpha(n) \geq \tfrac{1}{2}\psi_0, \tag{4.11}$$

and hence:

$$X_\alpha(n+1) \geq \begin{cases} X_\alpha(n) & \text{if } X_\alpha(n) \leq \psi_0, \\ \frac{1}{2}\psi_0 & \text{if } X_\alpha(n) \geq \psi_0. \end{cases} \tag{4.12}$$

We thus conclude that $X_\alpha(n) \geq \epsilon \equiv \min\{X_\alpha(1), \frac{1}{2}\psi_0\} > 0$, and our proof is complete. □

Under the assumptions of Lemma 4.3 above, Algorithm 2 remains well-defined for all $n \geq 0$ and the players' action choice probabilities never become arbitrarily

small.[15] With this in mind, the following theorem shows that Algorithm 2 converges to a connected set of QRE in potential games:

Theorem 4.4. *Let $\mathfrak{G}$ be a potential game and let θ be a penalty function with $x\theta''(x) \geq m$ for some $m > 0$. If the step size sequence γ_n of Algorithm 2 satisfies* (A1) *and the players' observed payoffs $\hat{u}_k$ are of the form* (4.5)*, Algorithm 2 converges* (*a.s.*) *to a connected set of QRE of $\mathfrak{G}$ with rationality level $\rho = 1/T$.*

Corollary 4.5. *With the same assumptions as above, Algorithm 2 with Gibbs updating given by* (4.8) *converges within $\varepsilon(T)$ of a Nash equilibrium; furthermore, the error $\varepsilon(T)$ vanishes as $T \to 0$.*

Proof. Proof of Theorem 4.4. Thanks to Lemma 4.3, Assumptions (A2) and (A3) for the iterates of Algorithm 2 are verified immediately – simply note that the innovation term of the strategy update step is bounded by the constant K of Lemma 4.3. Thus, by Lemma 4.1 and the subsequent discussion, $X(n)$ will converge to a connected set of *restricted* QRE of $\mathfrak{G}$. On the other hand, by Lemma 4.3, the algorithm's iterates will always lie in a compact set contained in the relative interior of $\mathcal{X}$, so any limit point of the algorithm will also be interior and our assertion follows. □

Remark 2. Importantly, Theorem 4.4 holds for any $T > 0$, so Algorithm 2 can be tuned to converge arbitrarily close to the game's Nash equilibria (see also the discussion following Theorem 3.8 in Section 3). In this way, Theorem 4.4 is different in scope than the convergence results of Cominetti et al. [9] and Bravo [7]: instead of taking high $T > 0$ to guarantee a unique QRE, players converge arbitrarily close to a Nash equilibrium by taking small $T > 0$.

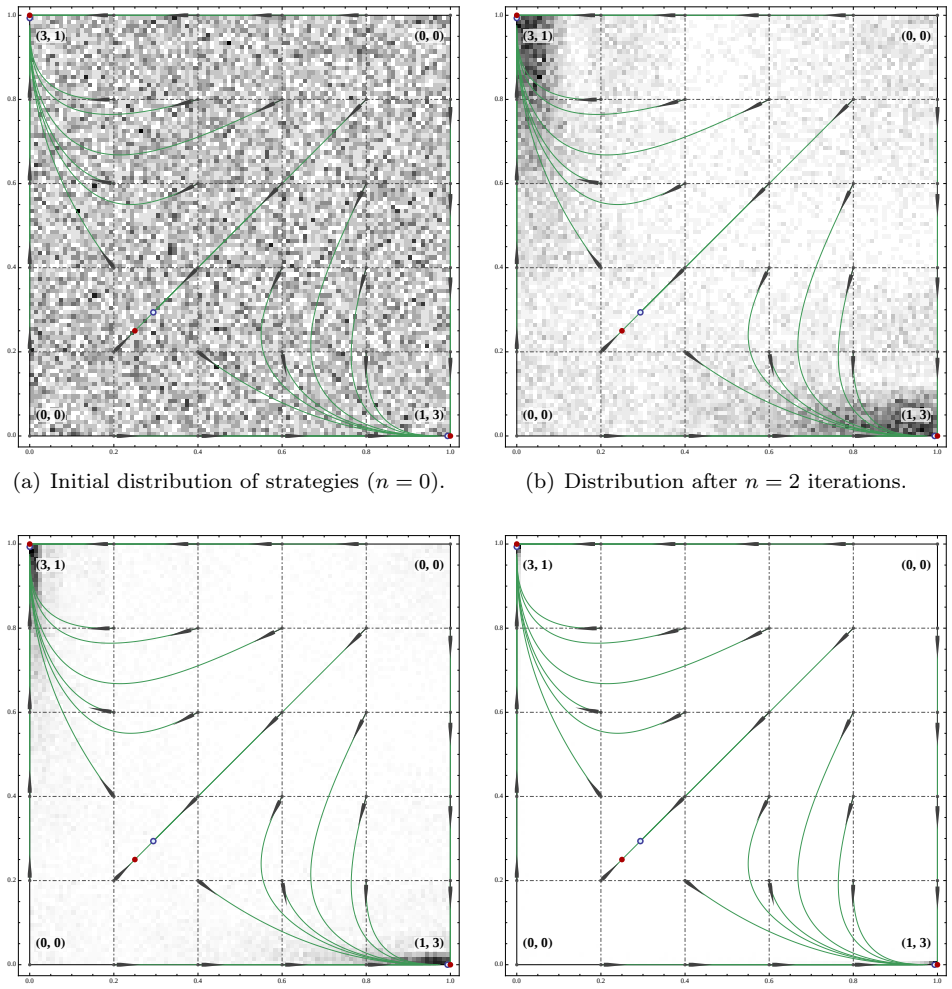
Remark 3. In view of the above, one might hope that Algorithm 2 converges to the game's (strict) Nash equilibria for $T = 0$. Unfortunately however, even in the simplest possible case of a single player game with two actions, Lamberton et al. [19] showed that the replicator update model (4.8) with $T = 0$ and step sizes of the form $\gamma_n = 1/n^r$, $0 < r < 1$, converges with positive probability to the game's globally suboptimal state.

4.4. **The robustness of strategy-based learning.** Even though Algorithm 2 only requires players to observe and record their in-game payoffs, it still hinges on the following assumptions:

(1) Players all update their strategies at the same time.
(2) There is no delay between playing and receiving payoffs.

Albeit relatively mild, these assumptions are often violated in practical scenarios: for instance, the decision of a wireless user to transmit or remain silent in a slotted ALOHA network is not synchronized between users, so updates and strategy revisions occur at different periods for each user – for a more detailed discussion, see

[15] In practice, it might not always be possible to obtain an absolute bound on the observed payoffs of the game (realized, estimated or otherwise). In that case, Lemma 4.3 cannot be applied directly, but Algorithm 2 can adapt dynamically to the magnitude of the game's payoffs by artificially projectings its iterates away from the boundary of the simplex – for a detailed account of this technique, see e.g. pp. 115–116 in Leslie [23].



(a) Initial distribution of strategies ($n = 0$).

(b) Distribution after $n = 2$ iterations.

(c) Distribution after $n = 5$ iterations.

(d) Distribution after $n = 10$ iterations.

FIGURE 3. Snapshots of the evolution of of Algorithm 2. In our simulations, we drew 10^4 random initial strategies in the potential game of Fig. 1 and, for each strategy allocation, we ran the Gibbs variant of Algorithm 2 with discount rate $T = 0.2$ and step sequence $\gamma_n = 1/(5 + n^{0.6})$. In each figure, the shades of gray represent the normalized density of states at each point of the game's strategy space; we also drew the phase portraits of the underlying mean dynamics (PD) for convenience. We see that Algorithm 2 converges to the game's QRE (which, for $T = 1/\rho = 0.2$ are very close to the game's strict equilibria) very fast: after only $n = 10$ iterations, more than 99% of the initial strategies have converged within $\varepsilon = 10^{-3}$ of the game's equilibria.

e.g. Altman et al. [1]. Furthermore, in the same scenario, message propagation delays often mean that the outcome of a user's choice does not depend on the choices of other users at the current timeslot, but on their previous choices.

In view of all this, we will devote the rest of this section to examining the robustness of Algorithm 2 in this more general, asynchronous setting. To that end,

let $R_n \subseteq 2^{\mathcal{N}}$ be the random set of players who update their strategies at the n-th iteration of the algorithm. Of course, since players are not aware of the global iteration counter n, they will only know the number of updates that they have carried out up to time n, as measured by the random variables $\phi_k(n) \equiv \text{card}\{m \leq n : k \in R_m\}$, $k \in \mathcal{N}$. Accordingly, the asynchronous variant of Algorithm 2 that we will consider consists of replacing the instruction "for each player $k \in \mathcal{N}$" by "for each player $k \in R_n$" and replacing "n" by "$\phi_k(n)$" in the step-size computation.

Another natural extension of Algorithm 2 consists of allowing the realized payoffs perceived by the players to be subject to delays (as well as stochastic perturbations). Formally, let $d_{j,k}(n)$ denote the (integer-valued) lag between player j and player k when k plays at stage n. Then, the observed payoff $\hat{u}_k(n+1)$ of player k at stage $n+1$ will depend on his opponents' past actions, and we will assume that

$$\mathbb{E}\left[\hat{u}_k(n+1) \mid \mathcal{F}_n\right] = u_k\left(X_1(n - d_{1,k}(n)), \dots, X_k(n), \dots, X_N(n - d_{N,k}(n))\right). \tag{4.13}$$

For instance, if the payoff that player k observes at stage n is of the form

$$\hat{u}_k(n) = u_k(\alpha_1(n - d_{1,k}(n)), \dots, \alpha_k(n), \dots, \alpha_N(n - d_{N,k}(n))) + \xi_k(n), \tag{4.14}$$

where $\xi_k(n)$ is a zero-mean perturbation process as in (4.5), it is easy to check that the more general condition (4.13) also holds.

In light of the above, we may derive a decentralized variant of Algorithm 2 as follows: first, assume that each player $k \in \mathcal{N}$ is equipped with a discrete event timer $\tau_k(n)$, $n \in \mathbb{N}$, representing the times at which player k wishes to update his strategies; assume further that $n/\tau_k(n) \geq c > 0$ for all $n \in \mathbb{N}$ so that player k keeps updating at a positive rate. Then, if t denotes a global counter that runs through the set of update times $\mathcal{T} = \bigcup_k \{\tau_k(n) : n \in \mathbb{N}\}$, the corresponding revision set at time $t \in \mathcal{T}$ will be $R_t = \{k : \tau_k(n) = t \text{ for some } n \in \mathbb{N}\}$. In this way, we obtain the following distributed implementation of Algorithm 2, stated for simplicity with logit action selection in mind (viz. $\theta(x) = x \log x$):

Algorithm 3 Strategy-based learning with asynchronous in-game observations

Parameters: $T > 0$, θ_k, γ_n
$n \leftarrow 0$;
Initialize $X_k \in \operatorname{rel\,int}(\mathcal{X}_k)$ as a mixed strategy with full support; # initialization
Repeat
  Event `Play` occurs at time $\tau_k(n+1) \in \mathcal{T}$;
  $n \leftarrow n + 1$;
  select new action α_k according to mixed strategy X_k; # current action
  observe $\hat{u}_k$; # receive realized payoff
  foreach action $\alpha \in \mathcal{A}_k$ **do**

$$X_{k\alpha} \leftarrow X_{k\alpha} + \gamma_n \left[\left(\mathbb{1}(\alpha_k = \alpha) - X_{k\alpha} \right) \cdot \hat{u}_k - T X_{k\alpha} \left(\log X_{k\alpha} - \textstyle\sum_\beta^k X_{k\beta} \log X_{k\beta} \right) \right]$$

update mixed strategy

  until termination criterion is reached

Following Chapter 7 of Borkar [6], we will make the following assumptions for Algorithm 3:

(1) The step sequence is of the form $\gamma_n = K/n$, where K is a positive constant small enough to guarantee that Algorithm 3 remains well-defined for all n (cf. Lemma 4.3).
(2) The strategy revision process R_n is a homogeneous ergodic Markov chain over $2^{\mathcal{N}}$; in particular, if μ is its (necessarily unique) stationary distribution, the asymptotic update rate of player k will be $\lambda_k = \sum_{A\subseteq\mathcal{N}} \mu(A)\,\mathbb{1}(k \in A) = \sum_{A\subseteq\mathcal{N}:k\in A} \mu(A)$.
(3) The delay processes $d_{j,k}(n)$ are bounded (a.s.): this condition ensures that delays become negligible as time steps are aggregated.

These assumptions can actually be weakened further at the expense of simplicity – for a more general treatment, see e.g. Borkar [6]. Still and all, we have:

Proposition 4.6. *Under the previous assumptions, the conclusions of Theorem 4.4 still hold for the iterates of Algorithm 3 with asynchronous updates and delayed payoffs.*

Proof. Proof. By Theorems 2 and 3 in Chap. 7 of Borkar [6], it is easy to see that Algorithm 3 represents a stochastic approximation of the rate-adjusted dynamics

$$\dot{x}_k = \lambda_k \mathrm{PD}(x_k), \tag{4.15}$$

where λ_k is the mean rate at which player k updates her strategy and $\mathrm{PD}(x_k)$ denotes the RHS of the rate-adjusted dynamics (PD). In general, the revision rate λ_k is time-dependent (leading to a non-autonomous dynamical system), but given that the revision process R_n is a homogeneous ergodic Markov chain, λ_k will be equal to the (constant) probability of including player k at the revision set R_n at the n-th iteration of the algorithm. These dynamics have the same rest points as (PD) and an easy calculation (cf. the proof of Lemma 3.3) shows that $F(x) = T\sum_k h_k(x_k) - U(x)$ is also Lyapunov for (4.15). The proof of Theorem 4.4 then goes through essentially unchanged. □

4.5. **Discussion.** We conclude this section by discussing some features of Algorithms 2 and 3:

- First, both algorithms are highly distributed. The information needed to update each player's strategies is the payoff of each player's chosen action, so there is no need to be able to assess the performance of alternate strategic choices (including monitoring other players' actions). Additionally, there is no need for player updates to be synchronized: as shown in Section 4.4, each player can update his strategies independently of others.
- The discount rate $T = -\log\lambda$ should be positive in order to guarantee convergence. Smaller values yield convergence to QRE that are very close to the game's Nash equilibria; on the other hand, such a choice also impacts convergence speed because the step sequence has to be taken commensurately small (for instance, note that the step-size bound of Lemma 4.3 is roughly proportional to the dynamics' discount rate). As such, tuning the discount rate T will usually require some problem-dependent rules of thumb; regardless, our numerical simulations suggest that Algorithm 2 converges within a few iterations even for small discount values (cf. Fig. 3).

References

[1] Altman, E., T. Boulogne, R. el Azouzi, T. Jiménez, and L. Wynter, 2006: A survey on networking games in telecommunications. *Computers and Operations Research*, **33 (2)**, 286–311.

[2] Alvarez, F., J. Bolte, and O. Brahic, 2004: Hessian Riemannian gradient flows in convex programming. *SIAM Journal on Control and Optimization*, **43 (2)**, 477–501.

[3] Benaïm, M., 1999: Dynamics of stochastic approximation algorithms. *Séminaire de probabilités de Strasbourg*, **33**.

[4] Björnerstedt, J. and J. W. Weibull, 1996: Nash equilibrium and evolution by imitation. *The Rational Foundations of Economic Behavior*, K. J. Arrow, E. Colombatto, M. Perlman, and C. Schmidt, Eds., St. Martin's Press, New York, NY, 155–181.

[5] Börgers, T. and R. Sarin, 1997: Learning through reinforcement and replicator dynamics. *Journal of Economic Theory*, **77**, 1–14.

[6] Borkar, V. S., 2008: *Stochastic approximation*. Cambridge University Press and Hindustan Book Agency.

[7] Bravo, M., 2011: An adjusted payoff-based procedure for normal form games, http://arxiv.org/pdf/1106.5596.pdf.

[8] Cabrales, A., 2000: Stochastic replicator dynamics. *International Economic Review*, **41 (2)**, 451–81.

[9] Cominetti, R., E. Melo, and S. Sorin, 2010: A payoff-based learning procedure and its application to traffic games. *Games and Economic Behavior*, **70**, 71–83.

[10] Fudenberg, D. and C. Harris, 1992: Evolutionary dynamics with aggregate shocks. *Journal of Economic Theory*, **57 (2)**, 420–441.

[11] Fudenberg, D. and D. K. Levine, 1998: *The Theory of Learning in Games*, Economic learning and social evolution, Vol. 2. The MIT Press, Cambridge, MA.

[12] Hart, S. and A. Mas-Colell, 2000: A simple adaptive procedure leading to correlated equilibrium. *Econometrica*, **68 (5)**, 1127–1150.

[13] Hart, S. and A. Mas-Colell, 2001: A reinforcement procedure leading to correlated equilibrium. *Economic Essays*, Springer-Verlag, Berlin, 181–200.

[14] Hofbauer, J. and W. H. Sandholm, 2002: On the global convergence of stochastic fictitious play. *Econometrica*, **70 (6)**, 2265–2294.

[15] Hofbauer, J. and K. Sigmund, 1998: *Evolutionary Games and Population Dynamics*. Cambridge University Press.

[16] Hofbauer, J., S. Sorin, and Y. Viossat, 2009: Time average replicator and best reply dynamics. *Mathematics of Operations Research*, **34 (2)**, 263–269.

[17] Hopkins, E., 2002: Two competing models of how people learn in games. *Econometrica*, **70 (6)**, 2141–2166.

[18] Hopkins, E. and M. Posch, 2004: Attainability of boundary points under reinforcement learning. *Games and Economic Behavior*, **53 (1)**, 110–125.

[19] Lamberton, D., G. Pagès, and P. Tarrès, 2004: When can the two-armed bandit algorithm be trusted? *The Annals of Applied Probability*, **14 (3)**, 1424–1454.

[20] Laraki, R. and P. Mertikopoulos, 2013: Higher order game dynamics. *Journal of Economic Theory*, **148 (6)**, 2666–2695.

[21] Laraki, R. and P. Mertikopoulos, 2013: Inertial game dynamics and applications to constrained optimization, http://arxiv.org/abs/1305.0967.

[22] Lee, J. M., 2003: *Introduction to Smooth Manifolds*. No. 218 in Graduate Texts in Mathematics, Springer-Verlag, New York, NY.

[23] Leslie, D. S., 2004: Reinforcement learning in games. Ph.D. thesis, University of Bristol.

[24] Leslie, D. S. and E. J. Collins, 2005: Individual Q-learning in normal form games. *SIAM Journal on Control and Optimization*, **44 (2)**, 495–514.

[25] Marsili, M., D. Challet, and R. Zecchina, 2000: Exact solution of a modified El Farol's bar problem: Efficiency and the role of market impact. *Physica A*, **280**, 522–553.

[26] McFadden, D. L., 1974: The measurement of urban travel demand. *Journal of Public Economics*, **3**, 303–328.

[27] McKelvey, R. D. and T. R. Palfrey, 1995: Quantal response equilibria for normal form games. *Games and Economic Behavior*, **10 (6)**, 6–38.

[28] Mertikopoulos, P., E. V. Belmega, and A. L. Moustakas, 2012: Matrix exponential learning: Distributed optimization in MIMO systems. *ISIT '12: Proceedings of the 2012 IEEE International Symposium on Information Theory*.
[29] Mertikopoulos, P. and A. L. Moustakas, 2010: The emergence of rational behavior in the presence of stochastic perturbations. *The Annals of Applied Probability*, **20 (4)**, 1359–1388.
[30] Monderer, D. and L. S. Shapley, 1996: Potential games. *Games and Economic Behavior*, **14 (1)**, 124 – 143.
[31] Ritzberger, K. and J. W. Weibull, 1995: Evolutionary selection in normal-form games. *Econometrica*, **63**, 1371–99.
[32] Rockafellar, R. T., 1970: *Convex Analysis*. Princeton University Press, Princeton, NJ.
[33] Rustichini, A., 1999: Optimal properties of stimulus-response learning models. *Games and Economic Behavior*, **29**, 230–244.
[34] Sandholm, W. H., 2010: *Population Games and Evolutionary Dynamics*. Economic learning and social evolution, MIT Press, Cambridge, MA.
[35] Sastry, P. S., V. V. Phansalkar, and M. A. L. Thathachar, 1994: Decentralized learning of Nash equilibria in multi-person stochastic games with incomplete information. *IEEE Trans. Syst., Man, Cybern.*, **24 (5)**, 769–777.
[36] Sharia, T., 2011: Truncated stochastic approximation with moving bounds: convergence. *ArXiv e-prints*.
[37] Sorin, S., 2009: Exponential weight algorithm in continuous time. *Mathematical Programming*, **116 (1)**, 513–528.
[38] Taylor, P. D. and L. B. Jonker, 1978: Evolutionary stable strategies and game dynamics. *Mathematical Biosciences*, **40 (1-2)**, 145–156.
[39] Tuyls, K., P. J. 't Hoen, and B. Vanschoenwinkel, 2006: An evolutionary dynamical analysis of multi-agent learning in iterated games. *Autonomous Agents and Multi-Agent Systems*, **12**, 115–153.
[40] van Damme, E., 1987: *Stability and perfection of Nash equilibria*. Springer-Verlag, Berlin.
[41] Weibull, J. W., 1995: *Evolutionary Game Theory*. MIT Press, Cambridge, MA.
[42] Young, H. P., 2009: Learning by trial and error. *Games and Economic Behavior*, **65 (2)**, 626–643.

Univ. of Versailles, PRISM, F-78035 Versailles, France
E-mail address: pierre.coucheney@uvsq.fr
URL: http://www.prism.uvsq.fr/users/pico/index.html

Inria, Univ. Grenoble Alpes, LIG, F-38000 Grenoble, France
E-mail address: bruno.gaujal@inria.fr
URL: http://mescal.imag.fr/membres/bruno.gaujal

CNRS (French National Center for Scientific Research), LIG, F-38000 Grenoble, France, Univ. Grenoble Alpes, LIG, F-38000 Grenoble, France
E-mail address: panayotis.mertikopoulos@imag.fr
URL: http://mescal.imag.fr/membres/panayotis.mertikopoulos